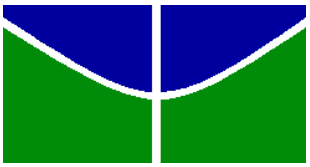

UNIVERSIDADE DE BRASÍLIA
Faculdade de Ciência da Informação
Programa de Pós-Graduação em Ciência da Informação

MARINA GARCIA DA SILVA PEREIRA

A APLICABILIDADE DO *BIG DATA* NAS PRÁTICAS ARQUIVÍSTICAS

Brasília - DF
2018

MARINA GARCIA DA SILVA PEREIRA

A APLICABILIDADE DO *BIG DATA* NAS PRÁTICAS ARQUIVÍSTICAS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Informação (PPGCInf) da Faculdade de Ciência da Informação (FCI) da Universidade de Brasília (UnB).

Área de Concentração: Gestão da Informação
Linha de Pesquisa: Organização da Informação
Orientadora: Profª Drª. Eliane Braga de Oliveira
Coorientador: Prof. Dr. Rogério Henrique de Araújo Júnior

Brasília - DF
2018

Pereira, Marina Garcia da Silva
GP436a A aplicabilidade do *Big Data* nas práticas arquivísticas / Marina Garcia da Silva Pereira; orientador Prof.ª Dr.ª Eliane Braga de Oliveira; coorientador Rogério Henrique de Araújo Júnior.
- Brasília, 2018.
86 p.

Dissertação (Mestrado – Mestrado em Ciência da Informação) -- Universidade de Brasília, 2018.

1. Documento de Arquivo. 2. Gestão de Documentos. 3. Arquivologia. 4. Big Data. 5. Dado. I. Braga de Oliveira, Eliane, orient. II. Henrique de Araújo Júnior, Rogério, coorient. III. Título

FOLHA DE APROVAÇÃO

Título: "A APLICABILIDADE DO *BIG DATA* NAS PRÁTICAS ARQUIVÍSTICAS"

Autor (a): Marina Garcia da Silva Pereira
Área de concentração: Gestão da Informação
Linha de pesquisa: Organização da Informação

Dissertação submetida à Comissão Examinadora designada pelo Colegiado do Programa de Pós-graduação em Ciência da Informação da Faculdade em Ciência da Informação da Universidade de Brasília como requisito parcial para obtenção do título de **Mestre** em Ciência da Informação.

Dissertação aprovada em 29 de junho de 2018.

Profª Drª Eliane Braga de Oliveira
Presidente (UnB/ PPGCINF)

Profª Drª Ivette Kafure Muñoz
Membro Interno (UnB/ PPGCINF)

Prof. Dr. Vanderlei Batista dos Santos
Membro Externo (Câmara dos Deputados)

Prof. Dr. Murilo Bastos da Cunha
Suplente – (UnB/ PPGCINF)

AGRADECIMENTOS

A Deus por guiar e iluminar meus caminhos.

Aos meus pais pelo apoio incondicional aos meus estudos, por proporcionarem a minha educação, base para minha vida.

Ao meu irmão, por me incentivar a entender visões diferentes de mundo.

Ao meu noivo Marcelo Alves Castro, por toda a paciência, debates entusiasmados e apoio na trajetória arquivística.

A minha orientadora Eliane Braga de Oliveira, por encarar comigo o desafio da investigação e da pesquisa, para aprofundar mais o meu conhecimento e colaborar com o fortalecimento da Arquivologia.

Ao professor Rogério Henrique de Araújo Júnior pelas contribuições que enriqueceram as discussões do tema deste trabalho.

Aos membros do Grupo de Pesquisa 'Fundamentos Históricos, Epistemológicos e Teóricos da Arquivologia – FHETA', em especial as professoras coordenadoras Cynthia Roncaglio e Angélica Alves da Cunha Marques, pelas discussões, novas abordagens da arquivologia e materiais enriquecedores para minha formação profissional.

A Vívian Miatelo, pela agilidade e presteza nas atividades da secretaria do PPGCInf.

Aos que foram meus estagiários no Confea (diretos ou não), em especial Iury Rodrigues e Guilherme Balduíno, por incentivarem meu ingresso no mestrado.

A equipe do projeto de *Big Data* no Secretaria de Inspeção do Trabalho (SIT), pela disponibilidade e contribuição a pesquisa.

A todos que me ajudaram durante o percurso da dissertação.

Querido Deus, Tu és minha proteção, a minha fortaleza. Tu és o meu Deus, eu confio em Ti. (Salmo 91:2)

RESUMO

Considerando a evolução tecnológica e seu impacto na gestão de documentos, no que é mantido nos arquivos e na recuperação da informação desejada, objetiva-se com esta pesquisa analisar as relações existentes entre as práticas arquivísticas e o documento de arquivo digital inseridos em um *Big Data*. Por meio da revisão de literatura são trazidos elementos para a análise dos temas como forma de compreender os fenômenos e explorar as suas interações. *Big Data* é a produção volumosa, veloz e diversificada de dados, documentos e informações digitais dos últimos anos somados à habilidade de coletar e analisar essas grandes quantidades de dados. De modo a corroborar este estudo observou-se um caso prático de um projeto de ferramenta para ecossistema de *Big Data* na Secretaria de Inspeção do Trabalho (SIT) do Ministério do Trabalho (MTB). Conclui-se que os documentos de arquivo serão utilizados como uma das fontes de dados a serem analisadas pelo *Big Data.* A gestão documental é relevante para a organização e recuperação da informação em tempos de *Big Data*, bem como o papel do arquivista mantém-se fundamental.

Palavras-chave: Documento de Arquivo. Gestão de Documentos. Arquivologia. *Big Data*. Dado.

ABSTRACT

Considering the technological evolution and its impact on the records management, on what is kept in the archives and on the retrieval of the desired information, this research aims to analyze the relationships between the archival practices and the digital record collected in a Big Data. Using the literature review method, elements are brought to the analysis of the themes as a way to understand the phenomena and explore their interactions. Big Data is the voluminous, fast and diversified production of data, documents and digital information observed in the recent years plus the ability to collect and analyze these large amounts of data. It was also analyzed a practical case study of a project for Big Data ecosystems in the Ministry of Labor, more specifically at Labor Inspection Secretariat as a subsidy for the decision-making process. It is understood that the records will be used as one of the data sources to be analyzed by Big Data. The records management remains relevant for the organization and retrieval of information in times of Big Data, as well as the role of the records manager and of the archivist remains fundamental.

Keywords: Records. Records Management. Archival Science. Big Data. Data.

LISTA DE QUADROS

LISTA DE SIGLAS E ABREVIATURAS

BI – *Business Intelligence*

Conarq – Conselho Nacional de Arquivos

DW – *Dataware House*

FGTS - Fundo de Garantia do Tempo de Serviço

ICA – *International Council of Archives*

InterPARES – *International Project on Permanent Authentic Records Electronic Systems*

ISO – *International Organization for Standardization*

LAI – Lei de Acesso à Informação

MTB - Ministério do Trabalho

OA – Open Access

OLAP - *Online Analytical Processing*

OLTP - *Online Transaction Processing*

OCR - *Optical Character Recognition*

SIGAD - Sistema Informatizado de gestão arquivística de documentos

SIT - Secretaria de Inspeção do Trabalho

SQL - Structured Query Language

TI – Tecnologia da Informação

UnB - Universidade de Brasília

VA – *Visual Analytics*

SUMÁRIO

INTRODUÇÃO

Jamais se produziu, se armazenou e se disseminou tanta informação como na sociedade atual. É fácil notar o quanto a tecnologia está inserida em nosso cotidiano, tanto no apoio a tarefas do dia a dia, no lazer, na velocidade da troca de informações, na comunicação, quanto na interação entre as pessoas.

A tecnologia avança e possibilita uma volumosa produção de dados e informações digitais, percebida ainda mais nos últimos anos com a popularização dos *smartphones*. Com a produção facilitada de imagens, de áudios e de vídeos na internet, temos mais testemunhas oculares e cidadãos-repórteres registrando a história e suas opiniões.

No âmbito profissional, o uso da tecnologia é um recurso indispensável seja na comunicação com os clientes, na edição e no envio de textos, nas transações financeiras ou no acompanhamento estratégico de metas e indicadores. Nas empresas, o registro dessas atividades e, portanto, do conhecimento produzido, é feito nos documentos de arquivo os quais tiveram sua produção simplificada com o uso da tecnologia. Isso permitiu um acúmulo maior e também mudança nos suportes, nos procedimentos e nos formatos de produção dos documentos. Esta crescente possibilidade de tecnologia gera impactos no que será tratado e mantido nos arquivos, acarretando novas exigências quanto aos métodos mais eficientes de busca e recuperação dos documentos.

Como responsáveis pela gestão e organização documental, os arquivistas precisam acompanhar a tecnologia e identificar alinhamentos entre ela e os requisitos arquivísticos, com o fim de realizar o tratamento técnico adequado dos documentos, promover e viabilizar o acesso a eles, preservar a sua organicidade e autenticidade, possibilitar a recuperação da informação, sempre atentos aos princípios e métodos arquivísticos.

A velocidade na produção de informações e de dados – estruturados ou não – em grandes quantidades e variedade, propicia uma nova forma de utilização das informações produzidas e acumuladas, é o chamado *Big Data*. Este examina, interpreta os dados produzidos e gera novos usos, o que tem impactado a condução de tomadas de decisão e a observação do comportamento dos usuários. Os

documentos de arquivo fazem parte deste cenário como mais uma das fontes de informação que servem de insumo para o *Big Data*.

Esta pesquisa identifica convergências e divergências na aplicação do *Big Data* nas práticas da Arquivologia, utilizando a revisão de literatura para análise e discussão de resultados, sob a ótica arquivística.

1.1 Justificativa

A motivação desta pesquisa decorre do interesse em investigar, na literatura científica, o impacto das inovações tecnológicas, em destaque, o *Big Data*, nas características dos documentos de arquivo e na prática arquivística.

No atual contexto tecnológico, verifica-se a necessidade de métodos e instrumentos que sejam capazes de fornecer soluções viáveis e econômicas para os problemas suscitados pela gestão de uma grande massa de dados e informações.

O conceito de *Big Data* advém da coleta e das formas de aproveitamento dos dados digitais, criados em grande volume, velocidade e variedade, nas várias fontes que os produzem. Seja para criar novos negócios ou gerar mais vendas, as empresas são as maiores utilizadoras de *Big Data*. Elas se utilizam da inteligência retirada dos dados para vantagem competitiva em seus negócios (MCAFEE; BRYNJOLFSSON, 2012).

Não apenas empresas, mas os governos começaram a utilizar o *Big Data* para analisar o volume de dados contidos em documentos que se repetem em seus órgãos como auxílio na elaboração futura de políticas públicas, obtendo informações estratégicas para investimentos eficientes, redução de gastos e melhoria na qualidade dos serviços prestados aos cidadãos.

Na busca por mais dados, a capacidade em obter as informações necessárias é crucial e a fonte principal onde podem ser encontradas está nos próprios arquivos das organizações e empresas que compõem-se dos documentos produzidos de forma natural e progressiva por uma instituição, pública ou privada, consequente às suas atividades para o cumprimento de suas finalidades, os quais são conservados de forma a servir de evidência, testemunho ou informação para quem os produziu ou para os demais cidadãos.

Questiona-se, então, como os documentos arquivísticos podem ser aproveitados pelo *Big Data*, enquanto fonte de informação? De que maneira essa

tecnologia irá se relacionar com os documentos de arquivo nas respectivas fases de seu ciclo vital? Ao serem retirados de seu contexto de produção, os documentos de arquivo servirão como fonte para o *Big Data*?

O momento atual com as possibilidades trazidas pela tecnologia, os documentos arquivísticos digitais e a recuperação da informação desejada impulsionam esta pesquisa a discorrer sobre a conceituação do *Big Data*, determinar suas características, retratar suas implicações e interfaces com a Arquivologia, de modo a identificar a pertinência das diretrizes adotadas no *Big Data* tendo em vista as práticas arquivísticas.

Dessa forma, pretendeu-se refletir sobre tal tecnologia à luz da teoria arquivística, expor novos recursos, identificar novos conhecimentos e cenários de atuação para o arquivista, o qual está habilitado a gerenciar grandes quantidades de documentos e informações orgânicas.

1.2 Problema

Tendo em vista os avanços tecnológicos e as mudanças na sua produção, de que forma os documentos de arquivo serão aproveitados pelo *Big Data*?

1.3 Objetivo Geral

Analisar a aplicabilidade do *Big Data* nas práticas arquivísticas.

1.4 Objetivos Específicos

i. Identificar as características e o funcionamento do *Big Data*;
ii. Verificar se o *Big Data* é aplicável aos arquivos e ao documento de arquivo, sendo capaz de preservar as características desses documentos;
iii. Analisar possíveis impactos dessa ferramenta no acesso e na gestão das informações assentadas nos documentos de arquivo.
iv. Analisar a aplicação do *Big Data* em uma realidade concreta – projeto em curso na Secretaria de Inspeção do Trabalho (SIT), no Ministério do Trabalho.

2 METODOLOGIA

Nesta pesquisa, a abordagem metodológica é qualitativa, pois ela não objetiva quantificar dados obtidos e/ou testar hipóteses, como, geralmente, faz a abordagem quantitativa, mas sim, se preocupa com a qualidade da informação na busca por entender razões e motivos para ações e interpretações. A pesquisa qualitativa foca em uma completa e detalhada descrição do que se observa, com contextualização, interpretação e entendimento de perspectivas (MACDONALD & HEADLAM, 2009).

Além do fato de possuir natureza teórica, esta pesquisa possui o propósito descritivo, com a finalidade de identificar as características de um problema e contemplar a descrição do comportamento dos fatos e fenômenos. De acordo com Bhattacherjee (2012), a pesquisa descritiva é dirigida a fazer observações cuidadosas e documentar detalhadamente um fenômeno de interesse.

Dessa forma, a revisão da literatura reforça as características apontadas acima, sendo a metodologia adotada nesta pesquisa, com a seleção de artigos de periódicos existentes sobre a temática, em bases de dados disponíveis com acesso *online*.

Em uma primeira aproximação com o tema escolhido para estudo, foi realizado um levantamento de referências teóricas em livros e na internet, partindo-se para a investigação de quantas e quais publicações e com quais conteúdos buscam uma inter-relação entre os campos do conhecimento da Arquivologia, Ciência da Informação, Tecnologia da Informação e mais especificamente, de *Big Data*.

Inúmeros e variados resultados foram obtidos e viu-se que o tema *Big Data* tem se popularizado nos últimos anos, principalmente após os anos de 2010 e 2011. Porém, poucos foram os resultados que indicaram relação entre à gestão documental ou com a informação orgânica produzida e registrada em documentos, decorrente das atividades organizacionais. O grande volume de pesquisas com o tema *Big Data* aborda, dentre outras, questões tecnológicas de funcionamento, *software* de análises, saúde (associado a temas cardíacos, estudos do câncer e sobre o genoma humano), previsão do tempo, aumento e estratégia de vendas, o que demonstra a pluralidade de aplicações do *Big Data* na sociedade atual.

A partir desses resultados múltiplos, partiu-se para a primeira etapa metodológica, na qual foram feitas buscas em bases de dados resultantes da primeira aproximação com o tema, durante três meses, utilizando como filtro palavras-chave em português e em inglês, visto ser este o principal idioma utilizado nas publicações internacionais. Na tentativa de alcance de resultados mais objetivos nesta análise, estabeleceu-se, então, o uso da palavra-chave: "*big data*". Cabe esclarecer que a utilização das aspas no termo de busca, serve para restringir a pesquisa aos materiais que contenham o tema de análise, excluindo o retorno de resultados que tenham as palavras "big" e "data" em separado, o que implicaria uma quantidade enorme de resultados de significados equivocados e dos mais variados temas.

As primeiras pesquisas foram realizadas nas bases: Google Acadêmico, *African Journals Online*, ANCIB - Informação & Tecnologia (Itec), arXiv.org, *Big Data & Society*, *Library and Information Science Abstracts/ LISA (ProQuest), Emerald Insight*, *Information, Communication & Society* e *International Journal of Digital Curation*. Inúmeras fontes e resultados foram novamente encontrados, porém sem que fosse possível identificar precisamente publicações que tratassem do documento arquivístico no contexto de *Big Data*.

Com mais restrições aos resultados, foram analisadas somente as publicações que contém o termo "*big data*" em seu título e cujo tema demonstrava a conceituação de *Big Data* e mais se aproximava do tema desta pesquisa. Foram encontrados não só artigos, mas livros, opiniões, apresentações, reportagens, resumos, teses e dissertações. A maioria das publicações está no idioma inglês e português, mas também foram encontradas em alemão, francês, espanhol, chinês, japonês e italiano. Todavia, mesmo com esta abordagem restritiva de busca, os resultados obtidos foram ainda em grande quantidade e variedade.

Dessa forma, uma nova estratégia foi empregada para a segunda etapa metodológica, restringindo-se às bases que indexavam apenas artigos de periódicos. Com esta seleção, foram excluídos outros veículos de publicação, e artigos que serviam de exemplo de aplicação do *Big Data* em outras áreas do conhecimento, tais como Medicina, Economia e abordagens técnicas da Computação.

Em continuidade à análise pretendida, foi estabelecido outro critério para a busca, qual seja selecionar as bases que indexavam periódicos da área de Ciência

da Informação e da Arquivologia, oportunidade em que foram selecionadas as bases: *Library and Information Science Abstracts/ LISA (ProQuest)* e *Emerald Insight*, por serem as bases internacionais que representam a produção científica das áreas de interesse.

Além disso, com o objetivo de obter artigos com a temática mais aproximada desta pesquisa, houve alteração na palavra-chave, agora definida da seguinte forma: "*Big Data*" AND (arquivologia OR arquivística OR "ciência da informação" OR "*information science*" OR "*archival science*" OR "*records management*").

Foram utilizados os filtros disponibilizados pela base *Library and Information Science Abstracts/ LISA (ProQuest)* que permitem escolher data de publicação, idioma e tema, de modo a sintetizar os resultados conforme os critérios da pesquisa. Dessa forma, foram selecionados os critérios: ano/período de publicação 2008 a 2017, os idiomas - português, inglês, espanhol e francês e, tema *big data*.

A base *Emerald Insight* não permite filtrar os resultados por data ou idioma, porém foi utilizado o filtro "*keyword*" *Big Data*. Sendo assim, obtivemos os resultados constantes no quadro 1.

Quadro 1 – Resultados iniciais da pesquisa bibliográfica

Base Consultada	Total de artigos encontrados
Library and Information Science Abstracts / LISA (ProQuest)	67
Emerald Insight	15
Total: 82	

Fonte: Elaboração própria.

Como mencionado, foram utilizadas as expressões ciência da informação e *information science* como palavras-chave, de forma a obter aproximações com o problema de pesquisa. Porém, os assuntos tratados nos artigos recuperados se restringiam às questões próprias da Ciência da Informação e não tratavam da interação ou relação possível entre a Ciência da Informação e o *Big Data* ou não estabeleciam interface com a Arquivologia. Diante disso, optou-se por excluir os resultados obtidos a partir das expressões ciência da informação e *information science* associados ao *Big Data*.

Foi assim necessário, portanto, realizar uma nova etapa para refinamento dos resultados, utilizando as palavras-chave: "*big data*" AND (arquivologia OR

arquivística OR "*archival science*" OR "*records management*") e o filtro assunto "*big data*". Foram obtidos oito artigos como resultado, compreendidos no período de 2014 a 2017, seis na base *Library and Information Science Abstracts/ LISA (ProQuest)* e dois na base *Emerald Insight*. Estes dois últimos artigos se repetiram como resultado e estão contidos nos seis primeiros resultados obtidos.

Assim, o *corpus* desta pesquisa é formado por seis artigos de periódicos, produzidos entre 2004 e 2017, com o tema *Big Data*, da base *Library and Information Science Abstracts/ LISA (ProQuest)*, os quais se encontram listados no Quadro 2.

Quadro 2 – Artigos resultantes da pesquisa bibliográfica

Autor(es)	Título	Ano de Publicação	Periódico de publicação
John McDonald; Valerie Léveillé	*Whither the retention schedule in the era of big data and open data?*	2014	*Records Management Journal*
Victoria Louise Lemieux; Brianna Gormly; Lyse Rowledge	*Meeting Big Data challenges with visual analytics.*	2014	*Records Management Journal*
Alan Rubel; Kyle M. L. Jones	*Student privacy in learning analytics: an information ethics perspective*	2016	*The Information Society*
Jens-Erik Mai	*Big data privacy: the datafication of personal information*	2016	*The Information Society*
Yanni Alexander Loukissas	*Taking Big Data apart: local readings of composite media collections*	2016	*Information, Communication & Society*
Johan Jarlbrink; Pelle Snickars	*Cultural heritage as digital noise: nineteenth century newspapers in the digital archive*	2017	*Journal of Documentation*

Fonte: elaboração própria.

Em decorrência do que foi apurado, podemos pressupor que este pequeno resultado reflete a pouca inserção do tema na Arquivística, se comparado aos demais materiais publicados com aplicações em outras áreas do conhecimento. Além dessa característica, nesta seleção, não foram encontrados artigos em outros idiomas, que não o inglês, mesmo com o filtro escolhido possibilitando resultados em português, francês e espanhol.

A continuação da pesquisa será feita com a análise dos artigos coletados para traçar, de forma argumentativa, as relações existentes entre a realidade observada e os conceitos estudados, bem como, identificar as possibilidades de

aplicação do *Big Data* na arquivologia, considerando as funções arquivísticas e as características dos documentos de arquivo.

3 Revisão de Literatura e Referencial Teórico

O ser humano se desenvolve a partir do convívio social, da interação com o ambiente e com os diversos elementos e ações que ocorrem ao longo da vida. Com a socialização, concebeu sociedades, estabeleceu meios de convivência, desenvolveu saberes, fazeres e tecnologias, criou valores, costumes e culturas.

Nesse cenário, a tecnologia evolui continuamente e, além de transformar a dinâmica da sociedade, amplia significativamente a comunicação, permitindo níveis de interação cada vez mais complexos, potencializando os processos comunicacionais e superando as limitações de representação de informações.

Inicialmente, a transmissão do conhecimento se dava fundamentalmente por meio da oralidade. Com a evolução das sociedades, exigiu-se uma forma perdurável de se fazer o registro das informações, de maneira que se mantivessem, ao longo do tempo, acessíveis às gerações futuras e como meio de preservar sua história.

O modo encontrado foi a escrita, cujo desenvolvimento considera-se fator essencial para a evolução da sociedade como a conhecemos hoje. Ao longo do tempo o suporte onde as informações eram registradas se transformou da pedra, da argila, do papiro ao papel, o qual, junto ao progresso da imprensa, permitiu maior produção e acúmulo de registros, os documentos.

O conjunto de vários documentos chama-se arquivo, como se pode observar na definição de Rousseau e Couture (1998):

> O conjunto das informações, qualquer que seja a sua data, natureza, ou suporte, organicamente [e automaticamente] reunidas por uma pessoa física ou moral, pública ou privada, para as próprias necessidades da sua existência e o exercício das suas funções, conservadas inicialmente pelo valor primário, ou seja, administrativo, legal, financeiro ou probatório, conservadas depois pelo valor secundário, isto é, de testemunho ou, mais simplesmente, de informação geral (ROUSSEAU; COUTURE, 1998, p. 284).

As instituições criadas para executar as tarefas de reunir, organizar, guardar e conservar os vários documentos foram também denominadas arquivo. Além disso, eram responsáveis por assegurar a legalidade e a autenticidade dos documentos testemunhais e de prova, que estavam sob sua responsabilidade.

Em consolidação a essas tarefas desenvolveu-se uma disciplina específica, a Arquivologia, a qual pode ser entendida como um conjunto de princípios, conceitos e técnicas a serem observados na produção, organização, guarda, preservação e uso de documentos de arquivo. Segundo os autores Rousseau e Couture (1998, p. 24) a

Arquivologia é “a disciplina que agrupa todos os princípios, normas e técnicas que regem as funções de gestão dos arquivos, tais como a criação, a avaliação, a aquisição, a classificação, a descrição, a comunicação e a conservação”.

Nesta pesquisa, considera-se que o objeto de estudo da Arquivologia são os documentos de arquivo - ou arquivísticos – conforme observamos nas seguintes definições:

Segundo Bellotto,

> Os documentos de arquivo são os produzidos por uma entidade pública ou privada ou por uma família ou pessoa no transcurso das funções que justificam sua existência como tal, guardando esses documentos relações orgânicas entre si. Surgem, pois, por motivos funcionais administrativos e legais. Tratam sobretudo de provar, de testemunhar alguma coisa (BELLOTTO, 2006, p. 37).

Para Rousseau e Couture,

> O documento é um conjunto constituído por um suporte [peça] e pela informação que ele contém, utilizáveis para efeitos de consulta ou como prova. (...) Documento de arquivo- documentos que contêm uma informação seja qual for a data, forma e suporte material, produzidos e recebidos por qualquer pessoa física ou moral, e por qualquer serviço ou organismo público ou privado, no exercício de sua atividade. Em resumo, um documento é constituído por um suporte ou peça e por um conteúdo (a informação nele registrada) (ROUSSEAU; COUTURE, 1998, p. 137).

Rondinelli afirma que,

> Documento arquivístico é a informação registrada, independente da forma ou suporte, produzida e recebida no decorrer da atividade de uma instituição ou pessoa e que possui conteúdo, contexto e estrutura suficientes para servir de prova dessa atividade (RONDINELLI, 2005, p. 129).

Sintetizando as definições temos que os documentos de arquivo são aqueles produzidos ou recebidos por instituição (ou pessoa) no transcurso das funções e atividades que justificam sua existência, guardando, esses documentos, relações orgânicas entre si. Servem como prova e testemunho das ações executadas pelo ente ou pessoa.

Consideramos, nesta pesquisa, principalmente os documentos de arquivo produzidos por instituições, cabendo o destaque quando considerarmos os documentos pessoais.

Conforme Lopes,

> O tratamento das informações contidas nos arquivos deve espelhar a vida, em especial, as vinculadas à razão de ser das organizações: as atividades fins. As de natureza meio, quase sempre são repetitivas e podem, mais facilmente, serem modeladas, também com o recurso da pesquisa (LOPES, 1997, p. 6).

O registro das atividades desenvolvidas e, consequentemente, o conhecimento institucional, estão nos documentos que irão atender necessidades informacionais que possam surgir.

De forma a manter as suas características e poder cumprir com suas funções, os documentos de arquivo precisam ser geridos adequadamente. Assim, decorrente dos avanços tecnológicos do período pós Segunda Guerra Mundial, que levaram a uma grande produção de documentos – dita explosão informacional - e a necessidade de recuperação ágil e precisa de informações, desenvolveu-se o conceito da gestão de documentos. Conforme aponta Jardim:

> Desde o desenvolvimento da Arquivologia como disciplina, a partir da segunda metade do século XIX, talvez nada a tenha revolucionado tanto quanto concepção teórica e os desdobramentos práticos da gestão ou a administração de documentos estabelecidos após a Segunda Guerra Mundial. (...) Segundo o historiador norte americano Lawrence Burnet, a gestão de documentos é uma operação arquivística "o processo de reduzir seletivamente a proporções manipuláveis a massa de documentos, que é característica da civilização moderna, de forma a conservar permanentemente os que têm um valor cultural futuro sem menosprezar a integridade substantiva da massa documental para efeitos de pesquisa" (JARDIM, 1987, p. 1).

Houve uma modificação na tradição dos arquivos voltados exclusivamente para servir à pesquisa histórica, caracterizada por um processo de aproximação com a Administração, na medida em que a gestão estabelece procedimentos e rotinas, visando a racionalização e a eficiência na criação, manutenção, uso e avaliação, para controle do grande volume de documentos arquivísticos armazenados.

Bartalo e Aparecida definem:

> Entende-se que a gestão documental ou gestão de documentos é o trabalho de assegurar que a informação arquivística seja administrada com economia e eficácia; que seja recuperada, de forma ágil e eficaz, subsidiando as ações das organizações e tornando mais confiável o processo de tomada de decisão e a preservação da história e da memória (BARTALO; APARECIDA, 2008, p. 84).

Os arquivos passaram a ser considerados como parte da base de conhecimento de uma organização, fato que interfere decisivamente na gestão da informação. Sousa e Araújo Júnior preconizam que:

> Parte dessa base de conhecimento, talvez a maior, esteja dentro da própria organização. Uma parcela é de informação/conhecimento registrado. Isso pode ser encontrado na biblioteca, no arquivo e nas bases de dados não institucionais. Outra parcela é o chamado conhecimento tácito, que quando for de alguma forma registrado tornar-se-á arquivo. Assim, podemos inferir que o arquivo compreende o principal estoque informacional da base de conhecimento da organização (SOUSA; ARAÚJO JÚNIOR, 2017, p. 48).

A percepção da importância da informação como recurso estratégico, ligado à produtividade, reduz excessos burocráticos, otimiza a tramitação, racionaliza o fluxo documental. Os estudos para compreender a dinâmica de produção e transferência – junto à necessidade de recuperação mais precisa – levou ao tratamento dos documentos produzidos, recebidos e daqueles já arquivados.

Estes precisavam ser organizados para ser encontrados em meio a grande quantidade do volume total e ter uma destinação conforme o valor e a frequência de uso que representam no conjunto documental. Constatou-se que não é possível guardar tudo e, por outro lado, não se pode eliminar todo o acervo, porque se perderia um patrimônio documental de valor imensurável.

Os documentos perdem seu valor primário[1] ao longo do tempo, de forma que não se justifica o custo de seu armazenamento e o trabalho intelectual e técnico envolvido na sua organização, além do necessário. Há também o risco da ineficiência na localização e na recuperação dos documentos diante do imenso volume armazenado. Nesse sentido, na gestão de documentos a melhor forma é a aplicação desses esforços nos documentos que adquirem o valor secundário e, por sua vez, investir em sua preservação, descrição e recuperação.

3.1 A tecnologia e os documentos de arquivo

O contínuo desenvolvimento tecnológico torna cada vez mais ágil o processo de produção, processamento, disseminação e armazenamento da informação, seja por parte das instituições, seja pelos indivíduos e traz à Arquivologia mais opções de automação, novos desafios aos arquivistas e à gestão dos documentos, sobretudo no que se refere à guarda de longo prazo.

No mundo atual de comunicação sem fronteiras, a Arquivologia não pode ficar alheia ao uso de tecnologias que, afinal, são grandes aliadas para a disseminação da informação. O trabalho nos arquivos precisa ser modernizado para continuar a proporcionar o desenvolvimento institucional e a prestação de informações à

[1] Os valores inerentes aos documentos públicos modernos são de duas categorias: valores primários, para a própria entidade onde se originam os documentos, e valores secundários, para outras entidades e utilizadores privados. Os documentos nascem do cumprimento dos objetivos para os quais um órgão foi criado – administrativos, fiscais, legais e executivos. Esses usos são, é lógico, de primeira importância. Mas os documentos oficiais são preservados em arquivos por apresentarem valores que persistirão por muito tempo ainda depois de cessado seu uso corrente e porque os seus valores serão de interesse para outros que não os utilizadores iniciais. (SCHELLENBERG, 2006, p. 180)

sociedade. Como opções para promover a modernização, cita-se a microfilmagem, a digitalização e os sistemas informatizados.

A microfilmagem, muito popular no Brasil nas décadas de 1960 a 1990 utiliza a captação das imagens de documentos por processo fotográfico. A legislação referente à microfilmagem dos documentos de arquivo prevê a validade jurídica de suas cópias, que podem substituir os originais, diminuindo assim o volume de documentos e, consequentemente, reduzindo o espaço físico a ser ocupado por eles.

Na dinâmica do desenvolvimento tecnológico, o uso de computadores como ferramenta de trabalho acarretou na mudança do principal suporte até então utilizado - o papel – demandando uma nova adaptação dos arquivos. A associação com a tecnologia e a computação para a criação de documentos de arquivo tornou mais ágil o processo de produção e mais dinâmicos o acesso e o uso dos documentos. O modelo de organização e administração da informação nos arquivos foi transformado ampliado com o advento do computador, tornando o armazenamento e a preservação preocupações ainda mais presentes.

A digitalização aparece como solução, pois permite o acesso simultâneo e remoto aos documentos. Entretanto, a digitalização é uma forma de mudar o suporte da informação e não de substituir integralmente o papel que foi produzido primeiro. Dessa forma, os registros públicos originais, ainda que digitalizados, devem ser preservados de acordo com o disposto na legislação em vigor. Diferentemente da microfilmagem, a digitalização não possui amparo legal quanto à validade jurídica em substituição ao papel.

A digitalização possibilita que se dinamize o acesso e a disseminação das informações entre funcionários de uma empresa, com a visualização instantânea das imagens de documentos que precisam ser consultados e administrados de forma rápida e organizada. Outras vantagens são a facilidade de acesso e de distribuição; redução de tempo das atividades que requerem a análise de documentos e a conservação do arquivo físico, já que evita o manuseio frequente.

Entretanto, dentre tantas vantagens para o acesso, Jarlbrink e Snickars (2017) abordam a necessidade de zelo no processamento dos documentos em papel que são digitalizados.

A princípio, a mudança de suporte pode parecer simples, porém, muitos aspectos devem ser levados em conta para que a cópia digitalizada resultante

possua as mesmas informações e características do documento original e este não tenha seu conteúdo transformado. Caso contrário, conforme discutido pelos autores, está-se criando novos documentos, com novas informações que nem existiam antes, o que chamam de "*digital noises*" ou "ruídos digitais":

> O ruído é produzido quando o software é usado para tornar as páginas legíveis e pesquisáveis por máquinas. Os processos são automatizados e a criação de ruído não é intencional - no entanto, como as ferramentas são programadas para funcionar de maneira específica, o resultado também não é intencional[2] (JARLBRINK; SNICKARS, 2017, p. 1239, tradução nossa).

Conflitos de interpretação no processamento variam desde a qualidade do documento em papel, seu estado de preservação, a tinta utilizada, a formatação, a diagramação do texto e até mesmo o idioma empregado nos documentos e o idioma nativo do software de processamento que irá analisar as imagens digitalizadas e fazer o reconhecimento do texto que usa a tecnologia *Optical Character Recognition (OCR[3])*. Para citar um exemplo, no caso estudado pelos autores, palavras dos jornais suecos do século XIX acabaram por conter o símbolo do Euro, a moeda oficial na União Europeia, criado apenas em 1998. As consequências disso para pesquisadores são inúmeras, pois criam ilusões e incorreções históricas, que com o passar dos anos podem se tornar verídicas, mas são apenas erros de processamento.

Jarlbrink e Snickars (2017, p. 1239) destacam ainda o impacto na cadeia da proveniência do documento original causado pelas etapas de processamento estabelecidas para a digitalização, as quais também impactam o resultado final. As instituições que conservam os documentos originais estão conscientes das consequências, apesar da urgência pelo acesso e divulgação de informações. No entanto, é preciso conhecer estas questões de processamento da informação na digitalização, que mesmo que não intencionais, são riscos que devem ser evitados.

Apesar dos problemas apontados, o desenvolvimento da gestão de documentos passa por um momento de transição, no qual ainda temos a produção de documentos em papel, porém com uma tendência cada vez maior de criação e armazenamento exclusivos em formato digital.

---

[2] Noise is produced when software is used to make pages machine readable and searchable. The processes are automated and the noise creation is not intentional – yet, since the tools are programmed to perform in a specific way the result is not unintentional either (JARLBRINK; SNICKARS, 2017, p. 1239).

[3] Reconhecimento Óptico de Caracteres, ou OCR, é uma tecnologia que permite converter a partir de um arquivo de imagem ou mapa de bits como papéis escaneados, escritos a mão, datilografados ou impressos em dados pesquisáveis e editáveis por um computador.

3.2 Os documentos de arquivo digitais

A realidade que se consolida é que além dos documentos em papel, os arquivos lidam com documentos híbridos[4] e digitais, os quais são parcela significativa da produção documental na chamada Sociedade da Informação, tendo em vista que são diretamente influenciados pelo uso das novas tecnologias de informação e comunicação. Mesmo assim, esses documentos devem ser alvo da gestão documental e necessitam de tratamento adequado.

Isto demanda o aprofundamento da discussão sobre a gestão de documentos arquivísticos digitais, definidos por Rondinelli:

> O conceito é formulado a partir da junção das ideias de documento, documento arquivístico e documento digital. (...) pode-se dizer que o documento arquivístico digital é um documento, isto é, "uma unidade indivisível da informação constituída por uma mensagem fixada num suporte (registrada), com uma sintática estável", "produzido e/ou recebido por uma pessoa física ou jurídica, no decorrer das suas atividades", "codificado em dígitos binários e interpretável por um sistema computacional", em suporte magnético, óptico ou outro (RONDINELLI, 2013, p. 234).

Innarelli (2008, p. 26) explica a estrutura do documento digital, que possui três elementos o *hardware* (físico), o *software* (lógico) e a informação (*bits*) armazenada em um suporte.

Sejam digitais ou não digitais, os documentos de arquivo para serem caracterizados como tal, possuem características únicas que os definem e que precisam ser mantidas ao longo do ciclo vital. Admite-se nesta pesquisa as definidas por Duranti,

> As características de imparcialidade, autenticidade, naturalidade, inter-relacionamento e unicidade tornam a análise dos registros documentais o método básico pelo qual se pode alcançar a compreensão do passado tanto imediato quanto histórico, seja com propósitos administrativos ou culturais. (DURANTI, 1994, p. 52).

As informações registradas nos documentos de arquivo são imparciais, pois retratam apenas a atividade que os gerou, uma vez que a sua qualidade de testemunho de fatos e prova de determinada atividade permite a reconstituição de realidades passadas. Por outro lado, os criadores dos documentos são parciais, pois visam seus próprios interesses.

---

[4] Os documentos ou dossiês/processos híbridos são formados por uma parte digital e outra convencional (e-ARQ Brasil, 2011).

Já a autenticidade, conforme definido em Rondinelli (2005, p. 66) por Macneil, é “[...] a capacidade de se provar que um documento arquivístico é o que diz ser. A autenticidade refere-se ao modo, à forma e ao status de transmissão do documento, às condições de sua preservação e custódia”.

A autenticidade diz respeito à geração de um documento e às qualidades que o legitimam para que ele possa exercer a plenitude de sua função administrativa, inclusive em termos legais (Lopez, 2010). Rondinelli (2011) completa que a autenticidade reside na garantia pelas entidades produtoras ou custodiadoras que os documentos sejam os mesmos desde o início, não sofreram nenhum processo de adulteração e, portanto, são autênticos.

Segundo o Estudo n.º 16 (2005) do *International Council of Archives* (ICA) a autenticidade é a permanência ao longo do tempo das características originais do documento de arquivo no que respeita ao contexto, estrutura e conteúdo.

Se não houver procedimentos adequados de segurança e de preservação, a confiabilidade, a autenticidade e o acesso desses documentos ficam ameaçados e, portanto, eles não terão mais valor como prova das atividades (ROCHA; SILVA, 2007). Para documentos digitais, manter a autenticidade durante as etapas de gestão é tão importante quanto durante a preservação em longo prazo.

Um elemento importante para a autenticidade dos documentos digitais é o metadado, definido pela Câmara Técnica de Documentos Eletrônicos (CTDE) do Conarq (2016) como “dados estruturados que descrevem e permitem encontrar, gerenciar, compreender e/ou preservar documentos arquivísticos ao longo do tempo”. Conforme Lopez,

> O metadado, portanto, é o registro fidedigno capaz de garantir a autenticidade de um documento eletrônico, o qual, nesse caso, confunde-se com sua informação. Não obstante, o metadado garante que o conteúdo informativo não seja desprovido dos dados contextuais da origem arquivística do ato administrativo que o produziu, além de garantir a permanência de seu valor probatório (LOPEZ, 2004, p. 71)

Além disso, os metadados auxiliam na recuperação da informação, através da indexação de assuntos que aumentam a precisão das buscas realizadas pelos usuários. Lemieux, Gormly e Rowledge (2014) mencionam em seu artigo a norma *ISO* 15489, que na sessão 7.2, demonstra a importância da captura dos metadados para permitir o uso ao longo prazo dos documentos e dos dados que eles contêm ou que deles podem ser extraídos.

McDonald e Léveillé (2014) chamam atenção para a qualidade dos metadados usados para descrição, facilitar o acesso e a recuperação, proteção, monitorar e controlar ações, preservação ao longo do tempo e para eliminação de dados. Embora os metadados e os esquemas de metadados usados sejam complexos, são recomendados ao gerenciamento de bases de dados e especificações de guarda e eliminação dos documentos[5] (MCDONALD e LÉVEILLÉ, 2014, p. 114, tradução nossa).

A terceira característica dos documentos de arquivo para Duranti (1994) é a naturalidade, ou cumulatividade, ela diz respeito a forma como os documentos são acumulados de maneira contínua, progressiva, consequente as atividades que os geram, um após o outro. Os documentos de arquivo não são coletados, são naturalmente produzidos e acumulados.

A unicidade se refere ao caráter do documento de arquivo ser único de acordo com o contexto no qual foi produzido. Como o documento de arquivo é intimamente ligado ao seu contexto de produção, ele advém de uma atividade geradora.

A última característica abordada por Duranti (1994) é o inter-relacionamento, a organicidade inerente aos documentos de arquivo. Rodrigues a define da seguinte forma:

> Se um arquivo é formado por um conjunto de documentos que se originam de ações articuladas em prol da missão de uma entidade, tem-se que ele resulta em um todo orgânico cujas partes são inter-relacionadas de modo a fornecer o sentido do conjunto (RODRIGUES, 2006, p. 109).

Como vemos Santos:

> É essa característica que justifica o fato do documento arquivístico precisar ser contextualizado a partir de seus vínculos com os demais documentos antes de qualquer análise sobre a sua guarda e ao seu significado para a instituição (SANTOS, 2015, p. 119).

Ainda, por Bellotto,

> Um documento arquivístico isolado do seu conjunto não faz sentido. Ele contém, portanto, não uma informação qualquer, mas a que é vinculada a uma vasta cadeia e é parte indissolúvel do seu meio genético de criação, vigência e uso. É a organicidade a grande característica dessa especificidade dos documentos de arquivo (BELLOTTO, 2014, p. 4).

[5] Although the metadata and metadata schema used in these initiatives were complex, it was recommended that they should be the tools used to ensure that retention and disposition specifications were recorded and accounted for (MCDONALD e LÉVEILLÉ, 2014, p. 114).

As relações administrativas orgânicas se refletem nos conjuntos documentais. A organicidade é a qualidade segundo a qual os arquivos espelham a estrutura, funções e atividades da entidade produtora/acumuladora em suas relações internas e externas. As informações orgânicas registradas dão origem aos arquivos.

Além das características acima, ao fazer uma revisão do conceito de documento de arquivo, Santos (2012) acrescenta o conteúdo estável e a forma fixa:

> Documento arquivístico é um conjunto de dados estruturados, apresentados em uma forma fixa, representando um conteúdo estável, produzido ou recebido por pessoa física ou jurídica (publica ou privada), no exercício de uma atividade, observando os requisitos normativos da atividade à qual está relacionado, e preservado como evidência da realização dessa atividade (SANTOS, 2012, p. 113).

A forma fixa, conceituada pelo Interpares (2012)[6], é a característica de um documento arquivístico que assegura que sua aparência ou apresentação documental permanece a mesma cada vez que o documento é manifestado, ou pode ser alterada segundo regras fixas (i.e., é dotado de variabilidade limitada).

O conteúdo estável se refere a característica do documento arquivístico que torna a informação e os dados nele contidos imutáveis e exige que eventuais mudanças sejam feitas por meio do acréscimo de atualizações ou nova versão (Interpares, 2007b).

Conforme Santos,

> (...) o advento das tecnologias de informação, a produção de documentos digitais e a constatação de que não chegariam a fase permanente aqueles que não fossem submetidos a procedimentos adequados de produção, uso, trâmite, conservação e avaliação, muito provavelmente, estariam perdidos – principalmente devido à fragilidade do suporte e a obsolescência de *hardware* e *software* – valorizam a gestão de documentos arquivísticos (SANTOS, 2008, p. 177).

Importante destacar que o tratamento dos documentos abarca sua produção, tramitação, até a destinação final. Este ciclo vital sistematiza a teoria das três idades dos arquivos de acordo com as fases corrente, intermediária e permanente. São etapas fundamentais do trabalho de qualquer arquivista e envolvem as funções arquivísticas, que se aplicam à organização e ao tratamento dos arquivos independente das idades dos documentos.

Os meios de armazenamento dos documentos digitais deverão protegê-los de acesso, uso, alteração, reprodução e destruição não autorizados. A norma *ISO*

[6] Glossário do TEAM BRAZIL.

15489 (2001, p. 7 apud LEMIEUX; GORMLY; ROWLEDGE, 2014, p. 127) qualifica o documento de arquivo como utilizável:

> Um documento utilizável é aquele que pode ser localizado, recuperado, apresentado e interpretado. Deve ser capaz de apresentação subsequente como diretamente conectado à atividade do negócio ou da transação que o produziu. As ligações contextuais dos documentos devem conter as informações necessárias para uma compreensão das transações que os criaram e usaram[7] (ISO, 2001, p. 7 apud LEMIEUX; GORMLY; ROWLEDGE, 2014, p. 127, tradução nossa).

Reforça-se então que nunca irá ser perdida a preocupação com a integridade, autenticidade, fidedignidade, organicidade, unicidade e confiabilidade.

As instituições devem estabelecer políticas de preservação e possuir infraestrutura organizacional, bem como requisitos, normas e procedimentos para assegurar que os documentos arquivísticos digitais permaneçam sempre acessíveis, compreensíveis, autênticos e íntegros. McDonald e Léveillé ratificam afirmando que:

> Para atender a esses múltiplos propósitos, os documentos devem ser capazes de se relacionar uns com os outros e apresentar as qualidades de confiabilidade, precisão e autenticidade durante o período de tempo em que são necessários[8] (McDonald e Léveillé, 2014, p. 110, tradução nossa).

Deste modo, temos como referência à proposta de Rousseau e Couture (1998), como também em Santos (2008), de sete funções arquivísticas: criação, avaliação, aquisição, conservação, classificação, descrição e difusão dos arquivos. Elas perpassam todas as idades documentais. Assim como nos documentos de arquivo não-digitais, os digitais estão sujeitos às mesmas funções.

A função criação se refere ao momento que o documento passa a existir para a instituição. É a sua elaboração e registro no sistema informatizado. O arquivista contribui nesta função com o estabelecimento de procedimentos para padronizar a produção dos documentos, seu acesso e preservação conforme as características do documento, pois há de se garantir a autenticidade e diminuir riscos da obsolescência tecnológica. Também é estabelecida a utilização de recursos como assinatura eletrônica, certificado digital e a opção por produção dos documentos em formatos abertos.

---

[7] A useable record is one that can be located, retrieved, presented and interpreted. It should be capable of subsequent presentation as directly connected to the business activity or transaction that produced it. The contextual linkages of records should carry the information needed for an understanding of the transactions that created and used them (ISO, 2001, p. 7 apud Lemieux; Gormly; Rowledge, 2014, p. 127).

[8] To serve these multiple purposes, records must be capable of being related to one another and present the qualities of reliability, accuracy and authenticity for the length of time they are require (McDonald e Léveillé, 2014, p. 110).

A função aquisição corresponde à transferência e recolhimento da documentação, ao recebimento dos documentos nos arquivos intermediário ou permanente. Tanto no seu recebimento quanto na transferência da custódia cabe ao arquivista definir as regras para assegurar que o acervo adquirido esteja íntegro, autêntico e que será armazenado nas mesmas condições, sem que sofra modificações por parte do novo custodiador.

A inserção de metadados é essencial no momento da produção e da aquisição de documentos arquivísticos digitais. Além disso, auxiliam a classificação, a avaliação, a preservação digital e a difusão.

Quanto a função classificação considera-se que,

> A classificação não se resume a atribuir números, códigos e subdivisões a atividades e documentos, e sim é um processo de organização intelectual em que as características, as informações e o contexto de cada documento tratado refletem, com maior ou menor grau de complexidade, as funções e atividades desenvolvidas por uma instituição, a vida de uma pessoa ou os fatos de uma cidade ou país (ALBUQUERQUE; MADIO, 2013, p. 10).

Os documentos digitais também devem ser classificados no momento de sua criação, de modo a evitar que a documentação fique desorganizada em meio digital, perca seu contexto de produção dentro da instituição ou seu valor associado a essas características.

Segundo Santos; Flores (2016):

> Ao abordar a classificação na ótica dos documentos digitais, percebe-se que é humanamente impossível reconstruir a organicidade de massas documentais acumuladas em virtude da complexidade do meio digital. Sem realizar esta função os documentos arquivísticos digitais tornam-se logicamente dispersos, de difícil localização. Logo, a ausência de procedimentos de classificação afetará diretamente nos procedimentos simultâneos e posteriores como avaliação, descrição, preservação e acesso (SANTOS; FLORES, 2016).

Depois de estudos específicos feitos pelos profissionais responsáveis pelo acervo e que conhecem a melhor maneira de agrupamento conforme a realização das atividades geradoras dos documentos, a classificação deverá ser feita dentro dos padrões estabelecidos, de modo a proporcionar a recuperação da informação. Importante destacar a busca por linguagens padronizadas para representar o assunto dos documentos e facilitar a sua localização.

A função avaliação surgiu como resposta ao crescimento exponencial de documentos de arquivo. Ela define se os documentos serão preservados ou eliminados e, ainda, seus prazos de guarda nas organizações. Como dito

anteriormente, entre a conservação e a eliminação total, a avaliação requer o processo técnico de análise dos valores dos documentos de arquivo.

As funções de classificação e avaliação possuem destaque entre as demais, visto que a atuação do arquivista determinará a destinação final dos documentos e influenciará diretamente na seleção do que será ou não mantido além do que constituirá a identidade e a memória organizacional e coletiva.

A função de conservação dos documentos constitui-se no conjunto de procedimentos que visa a manutenção de sua integridade física, a fim de desacelerar o processo natural de degradação e evitar que os documentos sejam comprometidos por acidentes, falha humana e desastres da natureza. Uma das finalidades do trabalho do arquivista é preservar os documentos de valor secundário para possibilitar o seu acesso por usuários que tenham direito ou interesse na informação ali registrada. Para os documentos digitais, mais conhecida como preservação digital, surgem preocupações a longo prazo devido a fragilidade física dos suportes, a obsolescência tecnológica e a vulnerabilidade do meio digital. É preciso garantir a autenticidade, a integridade e a confiabilidade da informação bem como o acesso continuado à ela.

A função descrição consiste na elaboração de instrumentos de pesquisa que auxiliem na identificação, rastreamento, localização e a uso das informações (BELLOTTO, 2006). Traduz-se em um conjunto de elementos facilitadores na recuperação e localização dos documentos, o que liga estes aos pesquisadores e demais usuários. Abrange a elaboração de instrumentos de pesquisa e meios de busca, por meio de termos específicos, palavras-chave, indexadores, dentre outros.

A função difusão dos arquivos está relacionada à divulgação dos documentos de valor secundário, à publicidade da informação, bem como a acessibilidade dos documentos, o que aproxima o arquivo e o usuário da informação. A organização das informações registradas é a base do acesso para um usuário.

A difusão e o acesso ficavam restritos a exposições e visitas orientadas ao acervo. Porém, com o apoio da divulgação *online* na internet e a digitalização dos acervos, são possíveis pesquisas mais céleres em qualquer lugar que haja conexão. Ainda as instituições de arquivo que podem divulgar amplamente suas atividades, serviços e documentação, de forma mais flexível. A difusão na internet proporciona uma nova dinâmica para a pesquisa documental, já que facilita o acesso às fontes primárias. No entanto, é comum que haja documentos com informações sigilosas.

Por isso, é preciso restringir o acesso a estas informações na forma que a legislação vigente determinar.

3.3 A gestão dos documentos de arquivo digitais

Nas instituições os documentos digitais também se acumulam e é preciso otimizar o armazenamento e o acesso a esses registros, que passam então a ser geridos em sistemas eletrônicos. Tais sistemas devem ser concebidos respeitando os requisitos arquivísticos e outras funcionalidades como o controle da produção de versões dos documentos arquivísticos; rigoroso cuidado e controle da segurança do armazenamento e do acesso; automação dos cuidados de guarda e destinação da massa documental; captura, identificação, pesquisa e recuperação dos documentos neles referenciados.

Lopez (2004) reforça as características do documento de arquivo:

> Se a informatização dos arquivos não levar em conta tais especificidades, está fadada, no máximo, a tornar-se somente um sistema de gerenciamento eletrônico de informações, que não será capaz de garantir as finalidades probatórias dos documentos de arquivo (LOPEZ, 2004, p. 70).

À nova realidade acrescenta-se o que afirma Jardim (1992):

> Os enormes problemas que ainda nos colocam a avaliação, recolhimento, processamento e guarda dos chamados documentos arquivísticos tradicionais não justificam, porém, negligenciarmos as novas questões resultantes do processo eletrônico de produção documental, sob pena de contribuirmos para ampliar ainda mais as dificuldades de preservação e acesso ao patrimônio arquivístico do país (JARDIM, 1992, p. 258).

Assim, permanece o fato de que toda e qualquer atividade ligada à gestão de documentos digitais (ou não) deve garantir as especificidades dos documentos de arquivo, além de respeitar os princípios e as funções arquivísticas, a fim de que a contextualização documental e a manutenção do valor probatório dela recorrente não se percam.

Do mesmo modo, o avanço do controle eletrônico, desde o trâmite até as atividades de transferência e recolhimento, é um fator fundamental para impedir a formação de novas massas documentais acumuladas (LOPEZ, 2004).

No Brasil, umas das especificações de requisitos a serem cumpridos pelas instituições produtoras/recebedoras de documentos e pelos desenvolvedores de sistemas, a fim de criar e tramitar de forma eletrônica dos documentos e garantir a sua confiabilidade, autenticidade e acessibilidade é o Modelo de Requisitos para

Sistemas Informatizados de Gestão Arquivística de Documentos (SIGAD) - e-ARQ Brasil, normatizado pela Resolução nº 25 do Conselho Nacional de Arquivos (Conarq). De acordo com esta publicação técnica, SIGAD é:

> Conjunto de procedimentos e operações técnicas característico do sistema de gestão arquivística de documentos, processado eletronicamente e aplicável em ambientes digitais ou híbridos, isto é, composto de documentos digitais e não digitais (Conarq, 2011).

O SIGAD possibilita controlar o ciclo de vida dos documentos arquivísticos. A gestão arquivística de documentos compreende a captura, a tramitação, a utilização e o arquivamento até a sua destinação final, isto é, eliminação ou recolhimento para guarda permanente. Se for desenvolvido e/ou adquirido em conformidade com os requisitos do e-ARQ Brasil (2011) possibilitará que os documentos permaneçam autênticos e acessíveis. Os documentos de guarda permanente, a serem recolhidos para instituições de preservação, deverão observar orientações específicas relativas à preservação digital, que não estão contempladas pelo e-ARQ Brasil.

Sabe-se que essas tecnologias e sistemas de informação devem ser o reflexo das atividades que são realizadas nos documentos em suporte tradicional; não são seus substitutos e nem excluem a organização e o controle prévio do acervo. As técnicas e funções arquivísticas a serem aplicadas aos documentos de arquivo continuam válidas para os documentos digitais.

Além da utilização de SIGAD, dentre os principais elementos que colaboram para a produção e manutenção de documentos arquivísticos autênticos, citam-se a adoção prévia de políticas e programas de gestão de documentos, a implantação de repositórios arquivísticos digitais confiáveis RDC-Arq, a adoção de esquemas de metadados de identificação e gestão previstos no e-ARQ Brasil (Conarq, 2011).

O sucesso do uso da tecnologia em arquivos dependerá fundamentalmente da implementação prévia de um programa de gestão arquivística de documentos na instituição, pois por meio dos procedimentos de gestão, de preservação e a aplicação dos pressupostos teórico-metodológicos preconizados pela Arquivologia garantirão a existência de sistemas informatizados idôneos e, assim, documentos arquivísticos confiáveis e autênticos.

Os sistemas de informação somados à automação voltada para os arquivos vieram contribuir para uma melhor recuperação da informação e acesso, bem como, para as atividades do arquivista, nas quais é fundamental a sua intervenção precoce na concepção e implementação de sistemas de arquivo, com o intuito de assegurar

que todos os documentos de arquivo digitais com valor secundário sejam preservados autênticos, fidedignos, inteligíveis e utilizáveis.

Aos arquivistas são exigidas competências nesses requisitos, em tecnologia, quanto ao domínio da legislação e, também, em conhecimentos de preservação digital ao desenvolverem ou implantarem um sistema de gerenciamento nas instituições. Os cuidados do arquivista com relação aos documentos digitais permanecem em todo o seu ciclo de vida – novos suportes, formatos de arquivo informatizado e tipos de armazenamento – afinal, o trabalho continua o mesmo, apenas mudou-se o meio no qual os documentos estão transitando.

Após a tramitação no Sistema Informatizado de Gestão Arquivística de Documentos (SIGAD), os documentos de arquivo devem ser recolhidos a um repositório que permita que suas características originais de documento de arquivo sejam mantidas. Este repositório se assemelha ao arquivo permanente onde as informações e os documentos estão disponíveis para acesso do público, podem servir de fonte histórica e são os novos armazenadores da memória da instituição produtora, como reflexo de suas atividades.

Para Santos e Flores (2016),

> O repositório digital deverá ser o ambiente autêntico para a preservação dos documentos arquivísticos, e por isto é reforçada a sua conformidade com o modelo OAIS. Além disso, é preciso realizar auditorias, tanto internas, quanto externas, a fim de comprovar a confiabilidade de seus procedimentos (SANTOS; FLORES, 2016).

Pode-se perceber que o ser humano busca em soluções externas, nos registros, nos sistemas, nos equipamentos de armazenamento de memória e nas memórias artificiais, a compensação para o não esquecer, devido à limitação da sua memória física. Como em toda profissão, novos desafios surgem tanto para os arquivistas quanto para a Arquivologia. Os sistemas informatizados com as características tratadas acima vão ser objeto de memória coletiva, nos quais o profissional da informação é o maior responsável pelo que vai existir e se perpetuar no futuro.

Charles M. Dollar (1994) afirma que as tecnologias de informação e comunicação estão conduzindo a uma nova era de “documentação”. Os arquivistas devem estar capacitados para gerenciar, organizar e assegurar a informação futura, sem abandonar os princípios e técnicas da área, seja nesse ambiente digital seja em uma próxima inovação tecnológica.

É imprescindível, portanto, compreender as influências das inovações tecnológicas na arquivística em seu objeto de estudo o documento de arquivo. Podem-se ser citados como exemplo de tecnologias atuais: as Redes Sociais, *Business Intelligence* (BI), Internet das Coisas, Impressões 3D, Realidade Virtual, *Big Data* e Computação na Nuvem.

3.4 Influências das inovações tecnológicas na Arquivística: o *Big Data*

Em nosso cotidiano estamos cercados pela tecnologia, cada vez mais conectados à internet e em nossas ações "*online*" são feitas trocas de dados com sistemas informatizados. Nestes, cada funcionalidade coleta e utiliza dados na sua operação e, assim, milhões de dados são criados todos os dias por todas as pessoas do mundo que trocam interações.

Para Semidão (2014, p. 185) dado é "elemento primário; isento de significação; número; símbolo; primeira percepção; elemento material; externo à mente; indício; insumo para informação; ligado à tecnologia computacional." E conforme Le Coadic (2004), dado em informática, é a representação convencional, codificada, de uma informação em uma forma que permita submetê-la a processamento eletrônico.

Os dados são capturados no que digitamos, compartilhamos, publicamos (postagens), nas pesquisas de buscadores, nos celulares (*smartphones*), nos computadores, nos *tablets*, entre outros dispositivos. Sensores e câmeras de vigilância capturam imagens e movimento; O GPS (*Global Positioning System*, ou, em português, Sistema de Posicionamento Global) sincroniza mapas, coleta a nossa localização e nos indica aonde ir. Seja nas redes sociais, no dia a dia no trabalho, nos e-mails, na pesquisa acadêmica, em fóruns, em *blogs*, lendo notícias, nos *sites*, realizando compras *online*, nos sistemas financeiros, todos produzem, coletam e trocam dados. A quantidade de informação digital está ficando cada vez mais vasta mais rápido. Além do volume extraordinário de registros gerados, nota-se a velocidade com a qual são criados, também a diversidade de origens que os criam e a variedade dos tipos de dados - de textos e documentos, a imagens, sons, planilhas e aplicações.

A produção ou coleta dos dados passa pelo estágio de "dataficação do mundo" (CUKIER; MAYER-SCHOENBERGER, 2013, p. 54). Os autores utilizam

esse termo para significar que documentos em papel, palavras, imagens, referências geográficas e, até interações, tudo pode ser digitalizado, colocado em um formato que pode ser quantificado na forma de dado e/ou diretamente coletados de nossas ações diárias no meio digital – de cliques em *sites*, tempo de visita, dados de navegação, entre outros – servindo de insumo para o *Big Data* processar e analisar. É um importante passo na busca da humanidade por quantificar e compreender o mundo, onde vários elementos que não podiam ser medidos, armazenados, analisados e compartilhados antes agora fazem parte de bancos de dados (CUKIER; MAYER-SCHOENBERGER, 2013, p. 11).

Outro aspecto é a instantaneidade na produção e na obtenção dos dados. Eles estão disponíveis mais rápido, a exemplo dos dados de cliques de navegação que podem ser obtidos pouco tempo após serem capturados e comentários das redes sociais podem ser observados em tempo real.

Esse cenário foi alcançado devido à evolução tecnológica de armazenamento e processamento das últimas décadas, o que permitiu o barateamento das mídias de leitura, gravação e armazenagem dos dados, além da conectividade do mundo globalizado. E a variedade de fontes de dados está relacionada às diversas possibilidades de equipamentos ou aplicações envolvidos na geração e na captura, provenientes, especialmente, da Web 2.0, além dos sistemas convencionais.

À medida que a expansão do universo digital avança, se torna mais complexo para as empresas processar, armazenar, gerir, assegurar e disponibilizar a informação nele. Isso tem impacto para quem trabalha no gerenciamento e na transformação dos dados em uma informação valiosa, em algo que possa ser apresentado, vendido ou utilizado. Furlan e Laurindo (2017, p. 92) afirmam que “a coleta e o armazenamento de dados têm crescido rapidamente, sendo que a capacidade dos softwares comuns está aquém da necessária para capturar, gerenciar e processar tais dados num período de tempo conveniente”.

Não é de hoje que a tecnologia da informação manipula dados utilizando várias ferramentas e linguagens, como por exemplo, os bancos de dados[9]

[9] Conhece-se como base de dados (ou *database*, de acordo com o termo inglês) o conjunto dos dados que pertencem a um mesmo contexto e que são armazenados sistematicamente para que possam ser usados no futuro.

relacionais “tradicionais” (*OLTP*[10]) e a linguagem padrão desses bancos *SQL*[11]. Do mesmo modo, há alguns anos já se sabe que o cruzamento de dois bancos de dados (duas ou mais variáveis) pode ajudar nos negócios correlacionando vendas (números e valores) com padrões de comportamento e consumo dos clientes.

A tendência tecnológica de cruzar vários sistemas, bancos de dados, fontes de informação, formatos variados e, associado a esse dilúvio de dados, ao novo paradigma da forma de processá-los e geri-los chamou-se de *Big Data*.

Além do uso da estatística e da probabilidade, percebeu-se que o volume crescente de dados poderia representar mais do que apenas uma grande quantidade.

A tecnologia envolvida no *Big Data* está relacionada com previsões, em aplicar a matemática e a estatística a enormes quantidades de dados a fim de prever situações. Por exemplo, em um censo sobre uma população, possuir a maior quantidade de dados possível ou a sua totalidade sobre determinada situação ou serviço, permite, por meio da estatística, obter resultados reais e não mais aproximações. O cruzamento de dados de diversos sistemas leva a melhora de serviços e a previsão de fenômenos mais rápido do que apenas com a observação comum. A grande quantidade é fonte também de oportunidades para criar valor com impacto no crescimento econômico, no fornecimento de serviços e na inovação.

Tendo em vista este contexto, podemos dizer que *Big Data* refere-se a grandes conjuntos de dados complexos, tanto estruturados quanto não estruturados, cujo processamento técnico tradicional e/ou algoritmos não são capazes de operar. Objetiva revelar padrões ocultos e levou a uma evolução de um paradigma de ciência orientado por modelos para um paradigma de ciência orientado por dados.[12] (TAYLOR-SAKYI, 2016, tradução nossa). Não há mais dificuldade em gerar informações e, sim, em analisar o conteúdo para orientar ações futuras de forma diferenciada.

---

[10] OLTP (*Online Transaction Processing* ou Processamento de Transações em Tempo Real) são sistemas que se encarregam de registrar todas as transações contidas em uma determinada operação organizacional.

[11] *Structured Query Language*, ou Linguagem de Consulta Estruturada ou **SQL**, é a linguagem de pesquisa declarativa padrão para banco de dados relacional (base de dados relacional).

[12] Big Data refers to large sets of complex data, both structured and unstructured which traditional processing techniques and/or algorithms are unable to operate on. It aims to reveal hidden patterns and has led to an evolution from a model-driven science paradigm into a data-driven science paradigm (TAYLOR-SAKYI, 2016).

As demais ferramentas de manipulação de dados tradicionais continuam a ser utilizadas pelas empresas quando é o caso ou também agora podem ser associadas para aprimorar o processamento para o volume do *Big Data*.

Para McDonald e Léveillé,

> *Big Data* pode ser definido como grandes quantidades e variedades de dados que, devido à sua disponibilidade rápida e às vezes "em tempo real", exigem extensa manipulação e mineração por meio da intervenção de várias tecnologias e ferramentas não tradicionais[13] (McDonald; Léveillé, 2014, p. 102, tradução nossa).

Conforme Pimenta, *Big Data*:

> (...) representa grosso modo o grande volume de dados, base para a produção de informações não estruturadas e estruturadas, produzidos de maneira exponencial na contemporaneidade. Mais do que seu volume, sua articulação em rede, sua velocidade e diversidade possibilitam a produção de mais dados, a partir dos dados já existentes, sobre indivíduos, grupos ou sobre a própria informação, quaisquer que seja ela, disponível (PIMENTA, 2013, p. 2).

Em definição pela International Data Corporation (IDC)[14]:

> *Big Data* é uma nova geração de tecnologias e arquiteturas, desenhadas de maneira econômica para extrair valor de grandes volumes de dados, provenientes de uma variedade de fontes, permitindo alta velocidade na captura, exploração e análise dos dados [15] (GANTZ; REINSEL, 2011, p. 6, tradução nossa).

Desses conceitos destacamos os primeiros três "V" característicos do *Big Data*: a variedade, o volume e a velocidade.

A variedade se refere aos muitos tipos, formatos (extensão do arquivo) e fontes de dados que se tem disponível. Como exemplo citam-se mensagens de e-mail, dados em cartões de crédito, em redes sociais, fotografias, áudios, vídeos e texto digitalizado. Para a gestão estratégica da informação, a seleção, a captação e o processamento desses dados devem estar alinhados em regras de negócio conforme a utilização que quer ser feita com os dados. Às empresas que conseguem captar essa variedade permite-se obter diferentes pontos de vista e realizar diferentes análises, agregando valor aos negócios.

---

[13] Big data can be defined as large quantities and varieties of data that, due to their fast and sometimes "real-time" availability, require extensive manipulation and mining through the intervention of various non-traditional technologies and tools (McDonald; Léveillé, 2014, p. 102).

[14] A IDC é principal empresa fornecedora mundial de inteligência de mercado, serviços de consultoria e eventos para os mercados de tecnologia da informação, telecomunicações e tecnologia de consumo (tradução nossa). Disponível em: https://www.idc.com/about.. Acessado em 16 de outubro de 2017.

[15] Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis (GANTZ; REINSEL, 2011, p. 6).

O volume se refere à imensa quantidade de dados gerados, grandes demais para serem processados e armazenados por *softwares* padrão, mas que proporciona mais opções de análises.

A velocidade remete tanto à produção de dados, na qual o fluxo de geração e transmissão de dados se torna tão elevada, que os sistemas tradicionais de análise não conseguem manipulá-los, quanto na necessidade de obter o que se deseja de forma instantânea ("*real time*") ou pelo menos mais rápida que os demais sistemas existentes. Além disso, a velocidade com a qual se obtém a informação determina a vantagem competitiva entre concorrentes.

Taurion (2015, p. 12) reforça os três "V's" e acrescenta o valor dos dados:

> *Big Data* trata não apenas da dimensão volume (...) mas existe também uma variedade imensa de dados, não estruturados, dentro e fora das empresas (coletados de mídias sociais, por exemplo), que precisam der validados (terem validade para serem usados) e tratados em velocidade adequada para terem valor para o negócio.

O valor informacional agregado após o tratamento e a gestão dos dados é o propósito norteador da utilização de uma ferramenta de *Big Data*, o que proporciona relevância aos dados obtidos. Estes permitirão descobrir melhores soluções aos problemas delimitados: se a empresa vai aumentar a sua receita, reduzir custos, identificar novas oportunidades, melhorar a qualidade do produto e/ou a satisfação do cliente. É necessário saber realizar as perguntas certas desde o início do processo da coleta até a análise de dados. A tendência a análise de dados promete transformar profundamente a maneira como vivemos, trabalhamos, fabricamos, valoramos e consumimos os produtos.

Uma das novidades trazidas pelo *Big Data* é a possibilidade do reuso dos dados para além de seu valor imediato, coletado inicialmente para um determinado fim. Os dados possuem valor para uso futuro podendo ser associados a outros dados em análises preditivas e na identificação de correlações que não tinham sido pensadas inicialmente na sua coleta. Assim, os dados podem ser reutilizados e adquirir mais valor.

Além da premissa do valor, temos o último "V", a veracidade dos dados, completando cinco "V" característicos do *Big Data*. A veracidade refere-se à confiabilidade dos dados, que devem possuir qualidade e consistência, ser fidedignos e íntegros aos dados armazenados e processados. Dessa maneira busca-se evitar a incerteza que pode surgir devido à inconsistência dos dados, o que

prejudicará o investimento e a análise da informação desejada. Por exemplo, os dados que são coletados internamente à instituição são mais confiáveis do que os coletados externamente, pois são produzidos pela própria instituição. Verificar os dados coletados quanto à adequação e relevância ao propósito da análise é decisivo para a obtenção de dados que agreguem valor.

Em síntese, temos no quadro 3 os cinco "V" característicos do *Big Data* e seus principais elementos:

Quadro 3 – Os 5 "V" do *Big Data*

Volume	**V**elocidade	**V**ariedade	**V**alor	**V**eracidade
-Terabytes; -Documentos; -Tabelas; - Distribuídos;	-Em tempo real (ou quase real); -Por lote; -Processos; -Em fluxo;	-Estruturados; -Não estruturados; -Multi-fator; -Probabilístico; -Vinculado; -Dinâmico;	-Estatístico; -Eventos; -Correlacionado; -Hipotético;	-Confiabilidade; -Autenticidade; -Origem, reputação; -Disponibilização; -Responsabilização.

Fonte: Adaptado e traduzido de Demchenko, Y.; Ngo, C.; de Laat, C.; Membrey, P.; Gordijenko, D. (2014).

De modo a relacionar os "V's" característicos do *Big Data* aos principais autores da literatura e sua descrição, apresentamos a síntese feita pelos autores Silveira; Marcolin; Freitas (2015, p. 48), conforme se observa no quadro 4.

Quadro 4 – Síntese das principais características do *Big Data*

CARACTERÍSTICAS	**AUTORES**	**DESCRIÇÃO**
Volume	(Demchenko *et al*., 2013)	Relacionado ao tamanho e quantidade de dados
	(Davenport, Barth & Bean, 2012)	Centenas de Terabytes ou Petabytes
	(Agrawal, 2014)	Grande quantidade e complexidade de dados
	(Mcafee & Brynjolfsson, 2012)	2,5 hexabytes de dados criados por dia, dobrados a cada 40 meses
Velocidade	(Goldman *et al*., 2012)	Necessidade de respostas em um curto prazo ou em tempo real
	(Demchenko *et al*., 2013)	Dinâmica de crescimento e processamento de dados
	(Mcafee & Brynjolfsson, 2012)	Dados capturados e processados quase em *real time*

	(Zikopoulos *et al*.,2013)	Velocidade de captura e análise de dados, formando um fluxo contínuo
Variedade	(Demchenko *et al*., 2013)	Diversidade de origens, formas e formatos de dados
	(Mcafee & Brynjolfsson, 2012)	Grande variedade de fontes e formas de dados com o desafio de encontrar os padrões de dados úteis para os negócios
Veracidade	(Demchenko et al., 2013)	Autenticidade, reputação de origem e confiabilidade dos dados
Valor	(Demchenko et al., 2013)	Dados com significado para os negócios, que contribuam com valor agregado

Fonte: SILVEIRA; MARCOLIN; FREITAS, 2015, p. 48.

Em seu artigo, Loukissas (2016), reporta-se a mais autores que estudam o *Big Data* e atrela cada um ao seu foco de pesquisa:

> (...) tecnicamente, em termos de "volume, velocidade e variedade" (Kitchin & McArdle, 2016); historicamente, como uma "mudança" de práticas passadas (Gantz & Reinsel, 2011); praticamente, com foco na dificuldade de gerenciá-los (Schneiderman, 2014); e ideologicamente, iluminando sua "mitologia" (Boyd & Crawford, 2012). Além disso, artigos populares sobre o Big Data os descrevem como ferramentas onipresentes para pesquisa e tomada de decisão em todos os domínios (Anderson, 2008; Lohr, 2012; Mayer-Schoenberger & Cukier, 2013)[16] (LOUKISSAS, 2016, p. 2, tradução nossa).

O objetivo de se aplicar uma ferramenta de *Big Data* é auxiliar as organizações a explorar o valor volumoso e repetido dos dados que podem estar nas bases de dados institucionais. Se antes, sem o poder de processamento de hoje, não conseguíamos obter todos, literalmente todos os dados estatísticos para uma análise, seja por custo, dificuldades de locomoção, ou tempo, hoje é possível que isso seja feito.

Outro fator é que também somos capazes de processá-los de várias maneiras a fim de obter novas associações e, assim, novos resultados não previstos. Como exemplo, temos a sua utilização pelas companhias Facebook e Twitter, nas quais a aplicação pode ser adotada para analisar padrões de compra de usuários em *sites* e sugerir novos produtos que eles possam ter interesse.

[16] (...) technically, in terms of their 'volume, velocity and variety' (Kitchin & McArdle, 2016); historically, as a 'step change' from past practices (Gantz & Reinsel, 2011); practically, focusing on the difficulty of managing them (Schneiderman, 2014); and ideologically, by illuminating their 'mythology' (Boyd & Crawford, 2012). Moreover, popular articles about Big Data depict them as ubiquitous tools for research and decision making across domains (Anderson, 2008; Lohr, 2012; Mayer-Schoenberger & Cukier, 2013) (LOUKISSAS, 2016, p. 2).

Big Data não se preocupa com a causa de um fenômeno, deixa os dados falarem por si. Tem a ver com a percepção e compreensão de relações entre informações que tínhamos dificuldade para entender (CUKIER; MAYER-SCHOENBERGER, 2013).

No contexto em discussão, não se trata de “usar o *Big Data*”. As ferramentas são para resolver “o problema” do *Big Data*. Nos últimos anos, a capacidade para “salvar” ou guardar dados se tornou algo relativamente simples e barato. Ou seja, montanhas de informação foram acumuladas em dispositivos de armazenamento. A dificuldade reside nos métodos de processamento, os quais não evoluíram com a mesma velocidade dificultando o acompanhamento do acúmulo de informações.

A proliferação dos dados e, consequentemente, da informação que representam, também a torna inacessível. É preciso, dessa maneira, organizar e dar sentido a todos esses dados e informações, pois a grande quantidade passa a influenciar a qualidade. Dessa maneira, não basta apenas produzir, coletar os dados e guardá-los. Saber se utilizar deles é fundamental para a eficiência e produtividade. Seu conteúdo e, consequentemente, a informação, precisam estar disponíveis no tempo certo e para a pessoa certa. Para tal, faz-se necessária uma organização e ela inicia com o processamento, o que representa um desafio nesse cenário de *Big Data*, uma vez que são diversas as fontes produtoras de dados.

Na definição trazida por Thomas Davenport passamos a explorar mais elementos:

> *Big data* é um termo genérico para dados que não podem ser contidos nos repositórios usuais; refere-se a dados volumosos demais para caber em um único servidor; não estruturados demais para se adequar a um banco de dados organizado em linhas e colunas; ou fluidos demais para serem armazenados em um *data warehouse* estático. Embora o termo enfatize seu tamanho, o aspecto mais complicado do *big data*, na verdade, envolve a sua falta de estrutura (DAVENPORT, 2017, p. 1).

Como visto acima os dados podem ser estruturados ou não estruturados, com base no seu gerenciamento e armazenamento. Os dados estruturados possuem esquema fixo, são organizados em linhas e colunas, geralmente são encontrados em banco de dados relacionais, representados em um formato estrito, são eficientes quanto à recuperação e processamento e são simples para relacionar informações. O tipo de banco de dados convencional utiliza *SQL* para organizar os dados. Porém, esses e outros bancos de dados relacionais populares não conseguem fornecer suporte simultâneo ao *Big Data* não estruturado, ao grande volume e a velocidade

dos dados devido a limitações de armazenamento, processos inflexíveis e expansão vertical, pelo motivo de que, quanto maior o servidor para processamento, mais caro ele é.

Já os dados não estruturados referem-se aos captados em formato não estruturado, isto é, não estão relacionados em tabelas de banco de dados convencionais, como vídeos, sons e sinais de geo-localização. Geralmente são dados de difícil recuperação e muitas vezes não dispõem de componentes necessários para identificação de tipo de processamento e interpretação, tornando o seu uso um desafio. Estima-se que 80% em média dos dados das empresas são não estruturados (REGO, 2013, p. 18).

E assim, no cenário de *Big Data* foi necessário estabelecer uma melhor forma para processar os "zilhões" de dados e essa forma foi utilizando o *NoSQL*[17], que são bancos de dados não relacionais, que usam diversos modelos de dados, incluindo documentos, gráficos, chave-valor e colunares. Possuem escalabilidade horizontal para aumentar o processamento e alta disponibilidade. Outra vantagem é o fato de ser possível a distribuição dos dados em servidores distintos, que permitem que se realizem cópias de segurança entre eles, a fim de manter o sistema funcionando mesmo que um dos servidores apresente algum problema. O intuito não é eliminar bancos de dados relacionais, mas oferecer uma alternativa para processar grandes volumes de dados.

Previamente ao *Big Data* outras soluções já permitiam uma maior manipulação dos dados. Temos como exemplo, na menção feita acima por Thomas Davenport, o *Data Warehouse* (DW), em português armazém ou depósito de dados, que são bancos de dados não relacionais, utilizam a tecnologia *OLAP*[18] e, armazenam dados corporativos detalhados. O DW possibilita a análise de grandes quantidades de dados coletados a partir de vários sistemas de negócio de uma empresa (fontes heterogêneas), voltado para aplicações de suporte à tomada de decisão. O DW analisa os dados, cria e organiza relatórios para ajudar as empresas na gestão das informações que fornecem competitividade e inteligência no mercado,

[17] *NoSQL* é um termo usado para descrever bancos de dados não relacionais de alto desempenho. Disponível em: https://aws.amazon.com/pt/nosql/. Acessado em 16 de outubro de 2017.

[18] *OLAP - On-Line Analytical Processing* (Processamento Analítico *Online*) é um *software* cuja tecnologia de construção permite aos analistas de negócios, gerentes e executivos analisar e visualizar dados corporativos de forma rápida, consistente e principalmente interativa. Disponível em: https://vivianeribeiro1.wordpress.com/2011/07/12/o-que-e-olap/. Acessado em 24 de outubro de 2017.

tornando os dados acessíveis a todos os usuários de níveis decisórios. Os dados são armazenados a partir de diversas fontes em uma estrutura estática e específica que define o tipo de análise que será feita nos dados já na fase de entrada.

O *Data Warehouse* tem sido a base para aplicações de *Business Intelligence* (BI). Seu conceito básico é um afunilamento ainda maior dos dados coletados do *Data Warehouse*, que chegam de forma exata e útil para a tomada de decisões. O BI transforma os dados brutos em informações úteis para analisar não só negócios como também as principais estratégias da entidade em questão.

Todavia, tanto o BI quanto o *Data Warehouse* são dependentes de dados estruturados. Porém, como vimos, a tendência é de crescimento em dados não estruturados. Assim, a estratégia utilizada pelo *Data Warehouse* não tem se mostrado suficiente para o *Big Data*.

Outro motivo é a necessidade da definição do tipo de análise a ser feita antes de armazenar os dados. Como características próprias do *Big Data*, as fontes agora são variadas, não são mais somente os bancos de dados internos das instituições, muitas questões podem ser ponto de partida de análises e, elas não são formuladas antes do armazenamento dos dados.

Tanto o *Data Warehouse* quanto o *Business Intelligence* continuam a ser usados, mas eventualmente com o aumento na quantidade de dados, no nível mais elevado de complexidade das análises estatísticas e a necessidade de informação em tempo real precisarão ser otimizados.

Todas as soluções são dependentes do dado de entrada. Percebe-se então que uma ferramenta para o *Big Data* será útil quando existirem grandes volumes de dados, estruturados e não estruturados e a melhor opção para o negócio depende de quais perguntas estão sendo feitas e quais os dados disponíveis, com relevância preponderante à qualidade das fontes de informação. A proposta de uma solução de *Big Data* é oferecer uma abordagem consistente no tratamento do constante crescimento e da complexidade dos dados, uma vez que eles não são alterados, são adicionados mais informações sobre eles.

O *Big Data* utiliza programas específicos, como o *Hadoop* e o *MapReduce*, para tratamento do alto volume de dados. O *Hadoop* é um projeto desenvolvido e mantido pela *Apache Software Foundation*, que consiste em um *framework* (família de projetos relacionados) de código aberto que utiliza infraestrutura de computação distribuída para processamento de dados em larga escala. Para garantir o

desempenho, o *Hadoop* faz uso do poder de processamento de sua arquitetura em *cluster* (grupo de servidores que funciona como um sistema único). Se necessário o *Hadoop* utilizará outras ferramentas auxiliares que o complementam formando um ecossistema *Hadoop: Sqoop, Spark, Zookeper, Oozie, Mahout, HBase, Pig, Flume* e outros.

O ecossistema *Hadoop* dispõe de dois serviços principais de armazenamento e processamento: o HDFS (sistema de arquivos distribuído - *Hadoop Distributed File System*) e o *MapReduce*, respectivamente. O *HDFS* manipula o armazenamento de dados entre todas as máquinas na qual o *cluster* do *Hadoop* está sendo executado. O segundo, o *MapReduce*, manipula a parte do processamento do *framework*.

O *MapReduce* é um mecanismo de processamento paralelo de dados, ele transforma dados maiores em menores, especificando em termos de funções de mapeamento e redução (duas fases de processamento: o *Map* e o *Reduce*). Ambas as tarefas rodam paralelamente no *cluster* e transformam os dados em menores quantidades de pares chave-valor. Assim, uma grande quantidade de dados pode ser processada e analisada, transformando-se em um montante menor de dados e com maior relevância. De modo a facilitar o entendimento, podemos pensar em uma grande tarefa que é dividida em várias tarefas pequenas, as quais são então executadas paralelamente em máquinas (*hardware*) diferentes e finalmente combinadas para chegar à solução da tarefa maior que deu início a tudo. O resultado final é realizado pela fase de redução e enviado para arquivos que conterão esses resultados. O escalonamento dos processos é feito internamente pelo *Hadoop*.

Além das ferramentas de processamento de dados temos aquelas responsáveis por auxiliar na análise, visualização e relatórios dos dados coletados, chamadas *data analytics*. São *softwares* que utilizam técnicas de recuperação de informação, reconhecimento de padrões e ferramentas de inteligência artificial - modelagem preditiva, mineração de dados (*Data Mining*), aprendizagem computacional (*Machine Learning*) – para permitir o tratamento estratégico das informações. Como exemplo, podemos citar o *Google Analytics,* utilizado para produção de relatórios de dados de navegação e interação com o objetivo de entender e otimizar o uso dos *sites* e páginas na internet. Também, citamos o *YouTube Analytics*, baseado na plataforma de vídeos, que permite monitorar a audiência e o engajamento dela com o conteúdo postado. Utiliza métricas e

relatórios quanto ao desempenho dos vídeos, do canal, origens de tráfego, informações demográficas e tempo de visualização.

Outra ferramenta para análise de dados é o *visual analytics* (VA), estudada no artigo de Lemieux, Gormly e Rowledge (2014). Ela emprega métodos para identificar padrões e tendências em grandes quantidades de dados e facilita a análise das informações de forma visual, muito útil ao *Big Data*.

O *visual analytics* se divide em três fases: coleta de dados e curadoria; pré-processamento de dados e análise, cuja estrutura provém organicamente dos dados em análise (LEMIEUX; GORMLY; ROWLEDGE, 2014, p. 124).

Na primeira fase, o processamento se inicia com a identificação, coleta e análise do tipo de dado disponível com o tipo de informação que o usuário espera extrair. Na etapa seguinte, é preciso mapear os dados, integrá-los, filtrá-los, agregá-los, modelá-los, momento no qual se pode transformar os dados em outra forma de visualização que seja mais fácil de ser analisada. Na terceira fase, os dados são apresentados como gráficos utilizando a ferramenta de *visual analytics*. Assim, quem vier a analisar os dados pode manipular e modificar sua posição, o que resultará em novos gráficos e novos sentidos aos dados.

Observamos dessa maneira que, por suas características e possibilidades de aplicação nas mais diversas áreas da atividade humana, o *Big Data* impõe desafios de ordem técnica, social, política e legal. É um novo ambiente, que requer novos entendimentos para coleta de dados e habilidades para serem manipulados.

3.5 Ética, Privacidade, Anonimidade

Como indicado anteriormente, *Big Data* se refere a estratégias e ferramentas para captura, armazenamento, análise e visualização de grandes volumes e formatos variados de dados. Ele é o resultado de uma sociedade em rede, é a plataforma para a qual os dados convergem e, muitas vezes nós mesmos oferecemos essas informações. Toda essa informação, quando captada, analisada e traduzida, revela perfis de comportamento e preferências de grupos de indivíduos. Isso proporciona oportunidades de negócio, contudo, também apresenta riscos na coleta, armazenamento e processamento de grandes conjuntos de dados.

Todos nós deixamos um rastro de dados, conscientes ou não, no uso da internet, seja nas buscas textuais, seja nas compras *online*, dentre outros. Há

também a coleta de dados pessoais nos cadastros de *sites* e sistemas, que são autorizadas após o consentimento do usuário, porém este não está atento ou não compreende os termos técnicos de tecnologia empregados nos termos de uso, o qual por sua vez, só permite prosseguir com ações após aceito. Segundo Mai (2016) essas informações pessoais são quaisquer dados que identifiquem uma pessoa específica.

Um número enorme de empresas rastreia o comportamento das pessoas *online*, geralmente através do uso de cookies[19] e, armazenam esses dados para usos posteriores, ou para melhoria dos serviços oferecidos, vendas e identificação de preferência de produtos. Outras empresas coletam dados com o objetivo de fazer ou vender propaganda direcionada a possíveis compradores ou vendem esses dados a terceiros. Mai afirma "As pessoas perceberam que as informações pessoais são, de fato, uma mercadoria que é vendida e negociada entre os impérios da informação e os corretores de dados"[20] (MAI, 2016, p. 192, tradução nossa).

Controlar a manipulação de dados pessoais, permitir ou bloquear acessos e a utilização desses dados, bem como, proteger a privacidade se torna um desafio, pois a informação se multiplica e é compartilhada em todo mundo. A proliferação de câmeras, fotos digitais e rastreamento de localização permite vigiar e controlar pessoas.

A forma de armazenar os dados também influencia a segurança e a privacidade de informações pessoais, já que uma base com informações centralizadas localmente e desconectada pode ser mais segura em relação a usos indevidos do que uma que esteja conectada e gerenciada por servidores de armazenamento fragmentados e espalhados pelo mundo.

Com essas tendências aumentam-se as responsabilidades na utilização das informações. Afinal, estão envolvidos dados pessoais, que trazem à tona a discussão por indivíduos e governos a respeito da ética e da privacidade na captura e uso dos dados no *Big Data*:

> O desafio ético contemporâneo na era do *big data* não é coletar informações pessoais. O fato é que as informações pessoais estão sendo

[19] Os **cookies são** utilizados pelos *sites* principalmente para identificar e armazenar informações sobre os visitantes. Disponível em: http://www1.folha.uol.com.br/folha/informatica/ult124u6772.shtml. Acessado em: 26 de outubro de 2017.

[20] "(…) people came to realize that personal information is in fact a commodity that is sold and traded among information empires and data brokers (MAI, 2016, p.192).

> coletadas e armazenadas por corporações privadas e órgãos públicos à medida que interagimos no ambiente digital. O desafio é determinar quando e como é eticamente responsável analisar as informações, o que procurar nos dados, quais perguntas fazer dos dados e a escala na qual é razoável fazer previsões sobre eventos futuros e ações baseadas sobre esses dados[21] (MAI, 2016, p. 194, tradução nossa).

Entre as questões que se colocam estão: a noção das pessoas de terem o direito ou a oportunidade de consentir autorização para fornecer ou quanto a utilização dos seus dados pessoais; a discussão da existência do direito moral do uso da informação se esta foi fornecida com consentimento pelo indivíduo.

Rubel e Jones (2016) exploram as questões de privacidade dos estudantes nas instituições de ensino superior no que tange à coleta de dados de navegação nos sistemas de informação do campus, conhecido como *learning analytics.* Desenvolvido em nome de melhorias, da eficácia e da efetividade institucional, o *learning analytics* serve para o aprendizado sobre o comportamento e os meios de aprendizado dos estudantes, cuja fonte para obtenção era somente por meio dos conteúdos e avaliações dos cursos. O desafio à privacidade que se aborda refere-se a necessidade da definição clara de um critério para justificar a coleta daquela informação que sustente a utilização do *learning analytics*.

Os benefícios devem superar os prejuízos à intimidade do aluno, pois mesmo que a informação seja útil há de se estabelecer normas e padrões com transparência para que o usuário tenha clara noção do que tipo de dado que disponibiliza ao usar os computadores do campus.

Também organizações e empresas esperam manter seus segredos e atividades e não desejam que sejam revelados a terceiros, para tal, muitas investem em segurança da informação.

Se por um lado, a partir da utilização dos dados é possível prever epidemias de saúde, por outro, com a obtenção da totalidade dos dados e, não mais uma amostra, pode-se facilmente descobrir por meio da comparação entre eles, a fonte produtora dos dados, seja uma empresa ou pessoa. Os riscos podem acarretar além de uma exposição da intimidade indevida, a supressão do direito ao anonimato e do direito ao esquecimento.

---

[21] The contemporary ethical challenge in the big data age is not whether to collect personal information. The fact is that personal information is being collected and stored by private corporations and public agencies as we interact in the digital environment. The challenge is to determine when and how it is ethically responsible to analyze the information, what to look for in the data, which questions to ask of the data, and the scale to which it is reasonable to make predictions about future events and actions based on that data (MAI, 2016, p.194).

O sigilo sobre a vida pessoal é imprescindível para evitar julgamentos, punições e demais prejuízos àquela pessoa ou grupo determinado, nos quais a situação pode evoluir para casos em que não haja possibilidade de defesa dos afetados. Entretanto, figuram no debate a liberdade de informação e a censura.

Dessa maneira, discussões são indispensáveis, com destaque ao aspecto legal, sobre a ética na coleta dos dados no *Big Data* e do uso de *softwares* estatísticos para análise de informações em termos do bem comum e sobre a vida pessoal do indivíduo. As soluções à questão são pensadas na forma de permitir aos usuários visualizar, analisar e atualizar políticas de dados de empresas, de acordo com suas necessidades e preferências, de forma a permitir a utilização ou não de seus dados pessoais.

3.6 A tomada de decisão com base em informações

Nas instituições várias decisões são tomadas pelos profissionais e por seus dirigentes diariamente. Elas podem ser simples ou complexas e guiam a gestão empresarial. Analisar as alternativas disponíveis e seus desdobramentos é um real desafio, principalmente porque uma boa ou má decisão afeta a competitividade e pode determinar a continuidade da empresa no mercado ou, no caso de uma má decisão nas entidades públicas, o desperdício de recursos públicos, o comprometimento de direitos e garantias, a inviabilização ou interrupção no andamento de políticas públicas, dentre outros prejuízos.

Um dos elementos fundamentais para a tomada de decisões é a informação que, correta e no momento certo, é crucial para a estratégia nos negócios. Conforme apontado Moreno: “é necessário que as informações sejam oportunas, relevantes, organizadas, disponibilizadas a fim de orientar os atores dos diferentes processos organizacionais na tomada de decisão” (MORENO, 2007, p. 14).

Utilizar a informação permite a vantagem competitiva no mercado, melhoria dos serviços e da produtividade, gerar uma economia de valor, maximiza oportunidades de marketing, embasa o desenvolvimento e o planejamento organizacional. McGee e Prusak (1994, p. 5) evidenciam que “a informação é capaz de criar valor significativo para as organizações, possibilitando a criação de novos produtos e serviços, e aperfeiçoando a qualidade do processo decisório em toda a organização”.

A ausência de informação clara e suficiente torna-se um obstáculo à tomada de decisão, pois as alternativas e riscos não estarão evidentes para a melhor escolha. Atualmente, a situação parece ser oposta dado o excesso de informação disponível, mas a identificação de quais informações são realmente relevantes dentre milhares de outras requer um processo sensível de diagnóstico de necessidades, o que gera alto valor agregado.

Como parte de sua estratégia, as instituições monitoram tanto o seu ambiente interno quanto o externo, analisam concorrentes, preveem riscos e oportunidades. Como esses ambientes estão em constante interação, conforme a necessidade, as informações serão buscadas tanto internamente quanto externamente à instituição. Desprezar um ou outro cenário pode criar obstáculos à estratégia organizacional e comprometer a qualidade de informação e a tomada de decisão.

O documento de arquivo é fonte essencial de informações internas para a tomada de decisão, já que é fruto do desenvolvimento das atividades realizadas pelas instituições. Por meio deles, dentre outros, se podem recuperar relatórios produzidos, informações financeiras, dados de empregados e reclamações de clientes. Os documentos de arquivo são utilizados quer pelo seu valor primário, para decidir, para agir e controlar as decisões e ações empreendidas, quer pelo seu valor secundário, para efetuar pesquisas retrospectivas que resgatam ações anteriores. Portanto, os documentos de arquivo transparecem as atividades executadas.

Para McGee e Prusak (1994) a criação, captação, organização, distribuição, interpretação e comercialização da informação são processos essenciais. Destas atividades as de criação, captação, organização e distribuição podem ser correlacionadas com as funções arquivísticas, as quais, com a gestão documental, promovem a efetiva recuperação dos documentos e a sua disponibilidade aos interessados, quando solicitados. Os documentos de arquivo proporcionam, ainda, fonte segura de informação, pois, para a sua utilidade estratégica, a informação requer integridade, precisão, fidedignidade, confiabilidade, autenticidade, qualidade e valor, características que lhes são inerentes.

Quanto melhor a qualidade das informações assentadas nos documentos, a celeridade na sua recuperação, a clareza na identificação dos problemas e o acesso às informações fundamentais possíveis de serem examinadas, melhor será a eficiência e a eficácia da decisão. Ela será melhor embasada proporcionando ao gestor menos riscos e efetivo exercício de poder.

Marchiori (2002, p. 73) ressalta que “as necessidades de informação se tornam cada vez mais complexas e dependentes de diferentes e múltiplas fontes – cuja correta avaliação e qualidade é fator crucial para os processos de tomada de decisão”. As informações também devem ser monitoradas em ambientes externos à instituição, os quais se modificam em velocidade e variedade e, requerem a ação de outras soluções tecnológicas.

A adoção de uma ferramenta de *Big Data* possibilitará a análise de dados tanto de origem externa quanto interna, estruturados e não-estruturados, otimizando-os para melhor recuperação de resultados e de uma quantidade viável de informações para ser manipulada, pois o volume disponível deve ser efetivo para a decisão.

O tratamento das informações, seja ela produzida interna ou externamente, constitui-se em um desafio. Porém, com o claro valor dos dados obtidos a informação é revestida de importância estratégica e a sua apresentação, por meio de gráficos e tabelas, possibilita a análise visual facilitada e correlação dos dados pelos gestores, atendendo às suas necessidades para a tomada de decisão.

3.7 Dados Abertos

Com base no artigo “*Whither the retention schedule in the era of Big Data and Open Data*” de John McDonald e Valerie Léveillé, incluído no resultado das pesquisas da revisão de literatura, será abordado o conceito de Dados Abertos.

Um projeto de dados abertos possui objetivos e públicos diferentes da utilização de uma ferramenta de *Big Data*. Este foca na obtenção de valor para os dados para utilização da própria instituição (uso interno), pública ou privada, enquanto os dados abertos focam no valor dos dados para utilização de um usuário externo à instituição pública detentora dos dados, geralmente em atendimento a políticas de transparência e publicidade dos atos públicos. Por exemplo, setores da indústria podem utilizar e reutilizar as informações divulgadas de órgãos governamentais para se desenvolverem social e economicamente.

Outra diferença é que, para o *Big Data*, novos métodos e sistemas são desenvolvidos para extração e manipulação de dados enquanto que para os dados abertos a tendência é de utilizar bases de dados menores já existentes ou estatísticas previamente configuradas e divulgadas no portal institucional. Neste

estágio há uma semelhança: a extração dos dados para os diferentes fins é retirada de sistemas baseados nos processos de negócio, nas informações internas disponíveis.

Os dados abertos também podem apoiar a tomada de decisão, pois reúnem informações institucionais que não necessariamente tenham sido consolidadas anteriormente. O apoio proporcionado se inicia nas etapas envolvidas na decisão de quais serão as bases disponibilizadas, como serão disseminadas e as forma de acesso.

Geralmente, as informações divulgadas obedecem a padrões mínimos de qualidade, de forma a facilitar o entendimento e a reutilização das informações. É preciso estabelecer as ferramentas de busca de conteúdo, formatar a informação combinando e associando estatísticas, configurar um portal de acesso para internet, monitorar acesso de usuários e seus *feedbacks*.

As características acima descritas de uma ação para dados abertos podem ser verificadas, como exemplo, no Brasil, com a Lei de Acesso à Informação (LAI) aprovada pela Lei nº 12.527 de 18 de novembro de 2011. Junto ao Decreto nº 7.724, de 16 de maio de 2012 que a regulamenta, estabelece a forma de divulgação e a periodicidade de atualização das ações do governo, determinando menus e conteúdos padronizados do que deve ser divulgado por todas as instituições governamentais em seus sítios eletrônicos. Além de cada página institucional, há o Portal da Transparência do Governo Federal com o objetivo de consolidar todas as informações e aumentar a transparência da gestão pública.

Para atender a legislação, os órgãos precisaram se organizar para reunir e consolidar as informações e então disponibilizá-las aos cidadãos. Neste contexto chama-se atenção para a importância da gestão documental, uma vez que sem que os documentos estejam organizados será ainda mais difícil divulgá-los. Além disso, a ausência de gestão documental pode comprometer o atendimento ao princípio da publicidade dos atos exarados pelo Estado. As instituições governamentais devem ainda seguir políticas de padronização de dados e formatos abertos de consulta, tais

como o Modelo de Acessibilidade em Governo Eletrônico (eMAG)[22] e os Padrões de Interoperabilidade de Governo Eletrônico (ePING)[23].

A LAI reforça o papel do cidadão de fiscalizador das contas públicas e permite solicitar o acesso a informações que não foram divulgadas, por meio dos Pedidos de acesso à informação ao Serviço Eletrônico de Informação ao Cidadão (e-SIC).

Outro exemplo de ação para disponibilização de dados abertos é a iniciativa do Portal Brasileiro de Dados Abertos, no qual mais de cem organizações[24] disponibilizaram acesso direto a bases de dados para consulta pelos próprios usuários. Referenciado por Oliveira (2015, p. 80), tem por objetivo ser a principal referência para publicação e reuso dos dados governamentais de órgãos do Poder Executivo Federal.

[22] BRASIL. Ministério do Planejamento, Orçamento e Gestão. Secretaria de Logística e Tecnologia da Informação. **e-MAG: Modelo de Acessibilidade em E-gov**. Ministério do Planejamento, Orçamento e Gestão, Secretaria de Logística e Tecnologia da Informação. 2014. Brasília: MP, SLTI, 92 p.

[23] BRASIL. Comitê Executivo de Governo Eletrônico. **e-PING – Padrões de Interoperabilidade de Governo Eletrônico** – Documento de Referência Versão 2017, Brasília, 41 p.

[24] Dados obtidos no sítio eletrônico do Portal Brasileiro de Dados Abertos. Disponível em: http://dados.gov.br/organization. Acessado em: 4 de abril de 2018.

4 PERSPECTIVA ARQUIVÍSTICA DO *BIG DATA*

Neste momento, passamos a analisar aspectos de convergência e divergência entre os objetos de estudo, os documentos de arquivo e o *Big Data*.

Pode-se considerar uma primeira diferenciação entre os objetos o fato de o *Big Data* referir-se a dados, enquanto a Arquivologia tem por objeto os documentos de arquivo. Entretanto, como será abordado a seguir, os documentos de arquivo podem servir de insumo, como mais uma fonte de dados, ao *Big Data*. Rousseau e Couture conceituam que:

> O termo dado é tirado da linguagem da gestão de informação, mas não deixa de constituir por isso uma realidade encontrada diariamente pelos arquivistas que têm de descrever e tratar dos documentos e, além disso, proceder à análise documental. O dado pode ser definido como a menor representação convencional e fundamental de uma informação (fato, noção, objeto, nome próprio, número, estatística, etc.) sob forma analógica ou digital, que permita efetuar o seu tratamento manual ou automático (informático) (ROUSSEAU E COUTURE, 1998, p. 137).

Dessa maneira, a partir dos documentos digitalizados ou dos nato-digitais, as informações assentadas nos documentos de arquivo podem ser exploradas como dados. A seguir, discutem-se requisitos necessários para que tal fato ocorra.

Com base no contexto em análise, sugere-se que o momento atual se configura na recorrência do fenômeno de explosão informacional, assim como ocorreu no período pós-Segunda Guerra Mundial. Um grande volume na produção de documentos e informações exigiu o desenvolvimento de novas tecnologias para o seu tratamento e automação. A gestão de documentos foi a solução adotada à época, juntamente a ferramentas tecnológicas e métodos para a recuperação da informação.

Atualmente, o avanço tecnológico, o acesso e a interação das pessoas com a tecnologia, a utilização do poder de processamento e a sociedade em rede, têm por consequência uma revolução de dados, cujos efeitos interferem em todos os cenários, do governo, aos negócios e à ciência. Novamente ocorre, então, a necessidade de desenvolvimento de novas tecnologias e métodos para lidar com o volume de informações.

O *big data* é uma dessas ferramentas. E possui em comum aos documentos de arquivo e, até com ações de dados abertos, o fato de estarem ligados e baseados nos processos de negócio da instituição na qual estão inseridos (MCDONALD; LÉVEILLÉ, 2014, p. 107). Desta maneira, o ponto de partida para

implementar novas ações e operações deve ser o entendimento dos processos de negócio, a forma como as funções desempenhadas por aquela instituição funcionam e se relacionam. São essas mesmas funções que agregam ao documento de arquivo a característica de organicidade, a qual é fundamental à gestão documental e é um dos princípios que guia a teoria e prática arquivística.

A confiabilidade dos dados a serem usados no *Big Data* depende da qualidade, integridade e completude dos controles que são definidos para os processos de negócio e que gerenciam a produção e a utilização dos dados. É preciso definir políticas de infraestrutura para a sua organização estabelecendo taxonomias, ontologias, linguagens, padrão de metadados, modelos de governança, definição de uso de *software*, maneiras de armazenamento, dentre outros, de forma a obter dados de qualidade, que possam ser cruzados com dados de diferentes fontes e com outros já existentes previamente, refletindo no potencial para serem explorados e serem feitas análises consistentes mesmo que coletados em grande quantidade e variedade no *Big Data*. Também deve ser possível rastrear os dados de volta à informação original das fontes das quais eles derivaram, no caso, com origem nos documentos de arquivo.

Isso somente acontecerá se os documentos de arquivo estiverem adequadamente inseridos em um programa de gestão de documentos, com ações referentes ao controle de sua produção, classificação, utilização, reprodução, acesso, arquivamento, armazenamento, eliminação, avaliação, destinação, de forma a manter íntegras as suas características intrínsecas, em um ambiente digital seguro, com gestão feita por sistemas computacionais que zelem por aquelas características, como um SIGAD e o repositório digital confiável.

Novamente, verifica-se que um arquivo ou um repositório digital, que não receba o tratamento arquivístico pertinente não permitirá o desenvolvimento de novas ações e iniciativas, pois não possibilita essa confiança. Além disso, inviabiliza a análise do conteúdo dos documentos, a recuperação de informações e a identificação do que possui valor ou não, o que também é relevante para o *Big Data*.

Como preconizado pela Arquivologia, o adequado tratamento dispensado aos documentos permitirá a sua recuperação, acesso e uso. A informação assentada nos documentos de arquivo deve estar disponível a quem se interessar. A classificação, a organização, os metadados e os padrões definidos para sua gestão

passam a ser ainda mais relevantes na produção digital de documentos, de modo a estruturar a informação e legitimar as relações hierárquicas e orgânicas entre eles.

Nesse sentido, os objetos estudados se assemelham na necessidade de organização, tanto dos documentos de arquivo quanto de outros dados coletados.

Corroboram a esta afirmação Araújo Jr. e Sousa (2016, p. 193): “É possível entender que a organização da informação deverá ser aplicada como subsídio para a definição do tratamento dos dados do *Big Data*, visando à recuperação da informação demandada”.

A organização é de tal forma fundamental, que torna o conteúdo dos documentos de arquivo mais uma fonte de coleta de dados para o *Big Data,* uma vez que manterá seu o caráter orgânico alinhado aos processos de negócio, bem como aos processos de aquisição e de agregação dos dados (MCDONALD; LÉVEILLÉ, 2014).

Em consequência à organização nos grandes conjuntos de dados das diferentes fontes de informação reunidas no *Big Data*, o contexto estará mantido, a relação orgânica entre os documentos de arquivo será preservada, o que permite sua identificação. Por fim, constituirão fonte íntegra e confiável de informação, contribuirão na análise dos demais dados coletados e permitirão a correlação e extração de relações antes não obtidas diretamente.

Neste sentido, Loukissas (2016, p. 6) afirma não existirem dados universais (*universal data*), ou seja, dados que sejam compreendidos em qualquer contexto. Este autor demonstra preocupação com a forma em que os demais estudiosos debatem o fenômeno do *Big Data* – sem considerar os diferentes locais de produção e uso dos dados – pois reconhece que os dados formam coleções heterogêneas, criadas a partir de vários locais de produção e modeladas com valores e normas diferentes, o que pode modificar pesquisas e práticas posteriores. Os dados estão inseridos no contexto que os produziu, na infraestrutura exigida para mantê-los, nos sistemas onde estão representados e no contexto social de quem o produziu[25] (LOUKISSAS, 2016, p. 6, tradução nossa).

Sem o contexto de produção, os documentos de arquivo perdem o sentido, perdem sua essência. Caso o *Big Data* não esteja baseado nos processos de negócio, os documentos de arquivo podem ser utilizados, apesar de que, com seu

[25] Data are situated within the means of their production, the infrastructure required to maintain them, their systems of representation, and the social order they reproduce (LOUKISSAS, 2016, p. 6).

vínculo rompido, serão apenas mais dados entre os vários outros utilizados pelo *Big Data*. Sem a confiabilidade e o valor da informação, as análises podem ser prejudicadas, o que pode dar origem a interpretações e correlações equivocadas entre os dados. Corre-se o risco, por exemplo, no entrecruzamento de dados feito por algoritmos que não levem em conta o contexto, de serem produzidos "*fake data*", isto é, dados distorcidos, falsos, fora do contexto real. São utilizados para propagar informações incorretas; atribuir soluções simples a questões complexas; embasar diagnósticos equivocados da realidade conferida a especialistas no assunto, ou causar danos na imagem de um indivíduo ou grupo. Isto reforça estereótipos e preconceitos. Saber interpretar, compreender e analisar resultados de forma crítica também é relevante.

Além de tudo, em um *Big Data*, no acúmulo dos dados de várias fontes, as diferentes práticas que produziram os dados podem aparecer e entrar em conflito: classificações diferentes, esquemas padronizados, erros de digitação ou digitalização, imagens borradas, erros tipográficos, diferentes infraestruturas que os mantêm (*hardware e software*), entre outros exemplos, marcam as várias práticas locais de tratamento dos acervos em cada período de tempo. Porém, para Loukissas (2016), ao contrário daqueles que tentam tratar, filtrar e solucionar esses "erros", essas questões não são prejudiciais, pois indicam traços importantes da produção daqueles dados, podendo indicar sua origem e contexto, constituindo-se em uma oportunidade sem precedentes para estudar a produção dos dados e levantar questões importantes sobre as histórias locais de culturas heterogêneas de coleta.

"Erros" na concepção dos dados foram também identificados por Lemieux, Gormly e Rowledge (2014, p. 127) apontados como prejuízos na reorganização e apresentação visual dos dados com o uso de ferramentas de *visual analytics* (VA):

> Indisponibilidade de dados; fragmentação de dados; qualidade de dados; valores faltantes; formato de dados; necessidade de padronização; modelagem de dados para compatibilidade técnica e melhor análise; desconexão entre a criação, o gerenciamento e o uso; necessidade de possuir esquemas para o controle de versões (..)[26] (LEMIEUX; GORMLY; ROWLEDGE, 2014, p. 127, tradução nossa).

Esses prejuízos foram comparados aos controles trazidos pelas normas *ISO* 15489 de 2001 - *International Records Management Standard* e *ARMA's Generally*

[26] Unavailability of data; Fragmentation of data; Data quality: Missing values, Data format, Need for standardization; Data shaping: For technical compatibility, For better analysis; Disconnect between creation/management and use; Recordkeeping: General expression of need for recordkeeping, Version control (LEMIEUX; GORMLY; ROWLEDGE, 2014, p. 127).

Accepted Recordkeeping Principles de 2009 (LEMIEUX; GORMLY; ROWLEDGE, 2014, p. 127). Podemos novamente notar que a organização aparece como necessária, pois esses prejuízos e uma melhor aplicação da ferramenta de *VA* seriam resolvidos com uma melhor gestão documental, pensada desde a origem da produção dos dados. Alguns exemplos de aplicações da ferramenta de *VA* na arquivística trazidos pelo artigo em questão:

> Lemieux e Baron (2011), por exemplo, observam as possibilidades de aplicação da análise visual para auxiliar os processos de descoberta eletrônica. Esteva e seus colegas discutem o uso da análise visual para a exploração de *e-mail* (Xu et al., 2011), e vários estudos examinam sua aplicação para as descobertas de achados arquivísticos, conforme pesquisado por Lemieux e Dang (2013). Kang et al. (2008) consideram o desenvolvimento de uma ferramenta VA para desduplicação de dados. Williamson aplica análise visual para investigar a qualidade dos metadados em um grande repositório digital (Williamson, 2013)[27] (LEMIEUX; GORMLY; ROWLEDGE, 2014, p. 126, tradução nossa).

Diante do exposto, pode-se perceber, segundo Araújo Jr. e Sousa (2016, p. 197), que os elementos trazidos pelo *Big Data*, entre eles, a coleta e armazenamento (organização), processamento, recuperação da informação, visualização dos dados e relatórios (representação), as técnicas de *machine learning*[28] e *data mining*[29], coincidem com atividades da gestão documental e com as etapas do ciclo documentário. Fica evidente que a concepção de uma ferramenta de *Big Data* deve considerar os princípios da política informacional e da gestão de documentos estabelecidas em uma organização.

Cabe esclarecer que o *Big Data* não deve ser confundido com um *software* para gestão dos documentos de arquivo, ou com um Sistema Informatizado de

---

[27] Lemieux and Baron (2011), for example, note the possibilities for applying visual analysis to aid e-discovery processes. Esteva and her colleagues discuss the use of visual analysis for email exploration (Xu et al., 2011), and several studies examine its application to archival findings aids as surveyed by Lemieux and Dang (2013). Kang et al. (2008) consider the development of a VA tool for data de-duplication. Williamson applies visual analysis to investigate the quality of metadata in a large digital repository (Williamson, 2013) (LEMIEUX; GORMLY; ROWLEDGE, 2014, p. 126).

[28] Um sistema de *Machine learning* (aprendizado de máquina) é capaz de analisar uma grande quantidade de dados por meio de métodos estatísticos específicos, além de usar uma variedade de algoritmos para encontrar padrões no banco de dados. Com base nesses padrões, consegue fazer determinações ou predições. Sua principal característica, porém, é não precisar ter as rotinas implantadas a mão: o próprio sistema tem a habilidade de aprender com a análise de dados e executar tarefa com uma precisão cada vez maior. Por Wesley Cleam. Disponível em: https://transformacaodigital.com/o-que-e-machine-learning-e-como-funciona/. Acessado em: 24 de julho de 2018.

[29] *Data Mining* consiste em um processo analítico projetado para explorar grandes quantidades de dados (tipicamente relacionados a negócios, mercado ou pesquisas científicas), na busca de padrões consistentes e/ou relacionamentos sistemáticos entre variáveis e, então, validá-los aplicando os padrões detectados a novos subconjuntos de dados. Disponível em: https://www.devmedia.com.br/conceitos-e-tecnicas-sobre-data-mining/19342. Acessado em: 24 de julho de 2018.

Gestão Arquivística de Documentos (SIGAD). Ele representa o fenômeno da grande produção de dados e de soluções técnicas desenvolvidas relativas ao uso e reuso para geração de novos dados.

O *Big Data* influencia as diferentes formas de produção e coleta de dados e se utiliza dos métodos da recuperação da informação, porém não faz gestão arquivística, arquivamento, preservação ou mantém a autenticidade. Os sistemas originários nos quais estão os documentos de arquivo continuarão com suas funções, o *Big Data* tem sua aplicabilidade na utilização, na difusão e no acesso aos documentos.

Devem ser mencionados, ainda, os documentos de arquivo produzidos por indivíduos, visto que eles podem ser coletados e também servir de insumo ao *Big Data*. Concomitantemente ao consumo de serviços e informações, os usuários produzem dados que são capturados, de forma consciente ou não, para ações de marketing, análise de comportamento e influência de conteúdo. Nossas “pegadas” ou rastros virtuais de ações em *sites* e redes sociais denunciam nossas preferências e podem servir para vigilância de nossas ações, conforme abordado no tópico sobre ética e privacidade no uso de dados pessoais.

Após compreender o que são os documentos de arquivo e o cenário que se define como *Big Data* pode-se concluir que a relação entre esses dois contextos é de que documentos de arquivo são insumos que servem ao *Big Data*, dentre as várias fontes utilizadas para análise e, vice-versa.

Os arquivos são fontes fidedignas de informações produzidas pela própria instituição e, sendo assim, são fontes confiáveis de dados, úteis ao processamento feito pelo *Big Data*. As instituições, a princípio, servem-se das informações orgânicas acumuladas no decorrer da execução de suas atividades e, somente quando essas não são suficientes para atender um objetivo desejado, buscam fontes externas de dados para obter as informações necessárias. A soma dessas várias fontes constituirá um ecossistema de *Big Data*.

Os documentos de arquivo auxiliarão ainda na complexa tarefa de validar os dados coletados das fontes externas. Além disso, a extração dos dados e sua análise pode levar a elaboração de novos documentos de arquivo decorrentes das atividades da instituição, que podem posteriormente servir também de insumo para o *Big Data*.

4.1 Ciclo vital, avaliação e valor dos documentos de arquivo

É possível perceber que nas práticas arquivísticas, a divisão entre as fases corrente e intermediária do ciclo vital se torna mais imprecisa com os documentos digitais. Se por um lado prevalece a produção daqueles com informações momentâneas, por outro, são constantemente demandados para embasar soluções e modelos já utilizados. Os documentos transitam entre as fases de forma mais volátil.

A fase intermediária, usualmente baseada nos custos com espaço físico, manutenção e frequência de uso, não se justifica mais por esses motivos. Não há a exigência para transferir os documentos para outra base de dados, ou servidor.

Mesmo com a baixa frequência de uso, os documentos não precisam ser separados em bases distintas e operam na mesma interface de acesso com as mesmas ferramentas de busca para sua recuperação. Tal mudança não impacta as rotinas feitas pelos usuários, mas pode dificultar a avaliação dos arquivistas na identificação do uso das informações, na contagem dos prazos de guarda e temporalidade e na verificação dos requisitos legais.

A possibilidade de se guardar tudo que se produz existe, porém ao longo do tempo verifica-se que esta não é a melhor solução devido ao custo de manutenção, pois os custos humanos, financeiros e materiais para salvaguardar os documentos digitais são maiores e se iniciam desde o planejamento de sua criação.

Diferentemente dos documentos em papel, não basta aguardar que os documentos digitais prescrevam armazenados em um local seguro. É preciso ter estratégias e ações para garantir que sua integridade seja mantida ao longo do tempo, com especificações para seu controle, monitoramento, armazenamento, gestão com responsabilidade e governança.

No que diz respeito à preservação digital, alguns aspectos a serem levados em consideração são: a maneira pela qual as mídias serão conservadas; custos com armazenamento em equipamentos de informática; migração no caso de mudança na tecnologia; obsolescência dos equipamentos de leitura e dos formatos de arquivo e, a manutenção adequada dos metadados para que o documento continue compreensível e acessível por o todo tempo necessário. Vimos esta aplicação também no contexto do *Big Data*, na afirmação de Sant’Ana:

> Quando se trata de preservar dados no contexto do *Big Data*, deve-se levar em conta não somente os aspectos comuns ao processo de preservação, mas, também, fatores, como por exemplo, a vasta gama de formatos e de variedades de fontes de dados, bem como a diversidade de dispositivos de coleta que, ainda, apresentam a constante evolução como agravante na questão da manutenção das informações sobre sua obtenção (SANT'ANA, 2016, p. 132).

Um dos riscos no caso de uma eliminação arbitrária é a perda das evidências e das trilhas documentais fornecidas pelos documentos, o que desconstitui a garantia de seu valor de prova e sua manutenção ao longo do tempo. Isto prejudicará a efetividade do *Big Data* que estava baseado no acúmulo de informações e no poder de retraçar o caminho das informações até a sua origem, se necessário.

Todos esses fatores são motivo para repensar e adaptar os instrumentos de controle de temporalidade e avaliação, sobretudo no cenário de *Big Data*. Os documentos de arquivo devem continuar a ser eliminados segundo os parâmetros estabelecidos na avaliação documental, de outra forma, os equipamentos de armazenamento esgotarão sua capacidade cada vez mais rápido e o volume de informações não relevantes será obstáculo na identificação do que é necessário ser preservado.

Para os autores McDonald e Léveillé (2014), todo o processo de elaboração dos instrumentos de guarda e destinação – a tabela de temporalidade[30] – deve se iniciar com a análise dos processos de negócio e, não, dos documentos. A elaboração deve iniciar na verificação da função de cada processo de negócio; do fluxo e a necessidade de cada tramitação – feita fisicamente ou em meio eletrônico; das regras que devem ser obedecidas por conta da legislação e estar associada à missão e ao contexto no qual a instituição está inserida. Deste modo, a avaliação permitirá identificar os documentos gerados, determinar seu valor probatório e/ou informacional, os requisitos para sua criação, captura, controle e especificações para guarda e eliminação de cada documento.

Como visto anteriormente, tanto os documentos de arquivo como os dados do *Big Data* estão atrelados à organização e à gestão documental, definidas a partir dos processos de negócio da instituição. Dessa maneira, como também ressaltado pelos autores, é preciso monitorar mudanças constantemente para que os requisitos

[30] McDonald e Léveillé (2014, p. 111) recomendam a utilização como guia para a elaboração de tabelas de temporalidade e destinação a ISO/TR 26122:2008 -*Technical report – Information and documentation – work process analysis for record.*

necessários à preservação das informações importantes sejam respeitados ao longo do tempo e que todos na instituição estejam conscientes da importância desses requisitos e das atividades pertencentes à gestão documental. Assim, nem a tabela de temporalidade nem o *Big Data* ficarão desatualizados ou com parâmetros dissonantes.

Mais uma vez, consequente a essas atividades da gestão documental, o valor da informação é mais um aspecto que a relaciona com o *Big Data*, o qual procura explorar diferentes usos no cruzamento dos dados coletados e nas análises preditivas.

A função avaliação com a análise dos processos de negócio e do potencial de uso da informação identifica os valores atribuídos (primário ou secundário) ao documento de arquivo. O valor primário refere-se ao uso administrativo, razão primeira da criação do documento, o que pressupõe o estabelecimento de prazos de guarda ou retenção anteriores à eliminação ou ao recolhimento para guarda permanente. Relaciona-se, portanto, ao período de utilidade do documento para o cumprimento dos fins administrativos, legais ou fiscais. Já o valor secundário refere-se ao uso para outros fins que não aqueles para os quais os documentos foram criados. Ele extrapola a função que motivou a sua produção inicial e detém valor probatório ou informativo, para fins de estudo ou pesquisa. Esses documentos de arquivo são guardados permanentemente a fim de garantir acessos futuros, com base no pressuposto de novos usos que podem vir a ter, além dos que possuem hoje.

Atividade já explorada pelos arquivistas por meio da função avaliação, Cukier e Mayer-Schoenberger (2013) afirmam “A evolução tecnológica proporcionou o aumento no poder de processamento de dados e informações e permitiu novas possibilidades no uso desses dados, indo além de seu uso primário previsto”.

Compreende-se que com o *Big Data*, a comparação entre a enorme quantidade dados, as possibilidades de visualização e estatísticas facilitaram prever e utilizar de formas diferentes os dados, possibilitando usos distintos daqueles que justificam sua criação, lhes proporcionando maior valor. Entretanto, tal ação não se constitui uma atitude nova na exploração da informação.

Novamente recorre-se aos autores McDonald e Léveillé,

> Os registros são a fonte de dados e informações valiosos que podem ser analisados para fins além do propósito por trás de sua criação ou coleta

> original - por exemplo, os dados derivados dos registros de todos os aplicativos de licença de dutos, quando combinados com dados ambientais, podem ser analisados para determinar quais terras do governo serão impactadas por possíveis dutos, onde os dutos podem ser localizados e onde os controles ambientais podem ser necessários (MCDONALD; LÉVEILLÉ, 2014, p. 110, tradução nossa)[31].

Assim sendo, temos que as informações que servirão de base para implantação de ferramentas de *Big Data* serão buscadas nos documentos de arquivo, serão obtidas por meio deles e a partir deles. Afinal, um dos motivos de sua gestão e organização é servir de alicerce e sustento para outros estudos, acesso e disseminação.

Frente ao exposto, sintetiza-se a relação entre o *Big Data* e a Arquivologia no quadro 5.

Quadro 5 – Quadro comparativo entre elementos do *Big Data* e da Arquivologia

Big Data	**Arquivologia**
Vínculo com processos de trabalho da instituição e com as funções institucionais.	Organicidade.
Podem utilizar dados de fontes externas e internas.	Os documentos de arquivo são registros das atividades da instituição.
Para confiabilidade do processamento de dados é necessário qualidade, integridade e completude dos controles do processo de negócio.	Os documentos precisam estar inseridos em: programas de gestão documental, SIGAD e repositórios digitais confiáveis.
Os dados precisam estar organizados.	Os documentos de arquivo devem ser classificados e organizados sob as regras da gestão documental.
Os dados precisam estar vinculados ao contexto de sua produção para fazerem pleno sentido.	Os documentos estão subordinados ao contexto de produção. Se utilizados no *Big Data* fora do contexto, tornam-se apenas mais um conjunto de dados, sem valores adicionais de autenticidade.
Objetiva a coleta de dados, difusão e acesso aos documentos.	Gestão arquivística, armazenamento, preservação, difusão, acesso, autenticidade.
Foco nos valores dos documentos para além dos objetivos de sua produção (valor secundário).	Foco nos valores primários e secundários.

Fonte: elaboração própria.

[31] Records are the source of valuable data and information **that can be analyzed for purposes beyond the purpose behind their original creation or collection** – for instance, the data derived from the records of all pipeline license applications, when combined with environmental data, can be analyzed to determine what government lands will be impacted by potential pipelines, where pipelines can be located and where environmental controls may be required (MCDONALD; LÉVEILLÉ, 2014, p. 110).

4.2 A utilização, pelo *Big Data*, de documentos e informações classificadas com grau de sigilo

No Brasil, em referência à Lei nº 12.527 de 18 de novembro de 2011, Lei de Acesso a Informação (LAI), parte-se do princípio de que as informações produzidas por entidades que compõem o Estado são públicas, ou seja, devem ser divulgadas e estar acessíveis para consulta. Mesmo assim, conforme explicita o artigo nº 23 da referida norma legal e seus incisos, existem aquelas informações que, por serem consideradas imprescindíveis à segurança da sociedade ou do Estado, devem ter acesso restrito, pois sua divulgação ou acesso irrestrito comprometeriam, por exemplo, a defesa e a soberania nacionais ou a integridade do território nacional.

Ao analisar o emprego do *Big Data* pelo Estado, utilizando-se, como fonte, documentos de arquivo dos vários órgãos que o compõe, sejam eles provenientes de arquivos correntes, intermediários ou permanentes, deve-se empregar um enfoque diferenciado àqueles documentos que contenham informações classificadas com grau de sigilo e que tenham restrição de acesso. Eles devem ser devidamente identificados na gestão documental, conforme impõem os requisitos de imprescindibilidade, para a segurança da sociedade e do próprio Estado e, também do acesso às informações de caráter pessoal.

No caso da utilização das informações com restrição ou sigilo por um projeto que utilize o *Big Data*, o uso deverá ser restrito às demandas internas do Estado ou do órgão executor do projeto, de forma a proteger as informações que são sigilosas do acesso indevido. Todavia, se o *Big Data* for empregado para estudos que serão divulgados além do âmbito do Estado, os documentos de arquivo que possuem restrição não devem ser utilizados como insumos, respeitando o grau de sigilo.

Neste contexto, os documentos em repositórios digitais confiáveis ou nos arquivos permanentes, que pressupõe acesso público visando atender o interesse de prestar informação à sociedade a respeito de um determinado tópico, são idealmente as fontes de documentos tratados e disponíveis que podem ser utilizadas como insumo ao *Big Data* sem maiores riscos de divulgação de informações com sigilo.

4.3 O papel do arquivista

Novas profissões surgem nesse cenário nos quais os dados estão amplamente disponíveis. Possuir a habilidade de explorá-los é cada mais requisitado e valorizado no estado atual da sociedade na qual vivemos. Um novo profissional, vinculado do *Big Data*, o cientista de dados, combina os conhecimentos de programador de *software* e de estatístico com as habilidades para extrair valor escondido nos dados. Porém, semelhanças são identificadas com as atividades e o conhecimento que um profissional já possui - o arquivista.

Um dos papéis da gestão documental é proporcionar às instituições a consciência das informações que produzem, como produzem, como armazenam e, também, entregar a informação precisa para tomada de decisão. Para que isso ocorra, deve-se possuir amplo conhecimento das funções organizacionais e das demandas por informações, aspectos inerentes à profissão de arquivista. Este deve ainda possuir habilidades no uso da tecnologia, *softwares*, técnicas de automação e recuperação da informação que mantenham as características dos documentos de arquivo, os preservem e mantenham os requisitos que garantam a presunção de sua autenticidade.

Destaca-se aqui a importância da capacitação após a formação regular e especializações em outras áreas relacionadas para enriquecimentos na prática e nas pesquisas na área arquivística.

As novas possibilidades delineadas a partir do volume, variedade e velocidade de acesso aos dados, trazem ao arquivista o desafio de planejar junto aos profissionais da ciência da computação e da informação as soluções informatizadas para gestão e repositórios de armazenamentos dos documentos de arquivo.

O arquivista irá contribuir com a gênese de sistemas e suas funcionalidades, pois conhece os procedimentos e processos fundamentais à gestão documental; define os metadados necessários na produção de cada documento; a necessidade informacional e a linguagem adequada aos usuários; as características e resultados esperados com a solução; o sigilo necessário às informações registradas nos documentos; o valor, a periodicidade, a manutenção dos documentos para usos imediatos e futuros; a proveniência das informações e a importância de se manter o contexto da produção do documento. Em resumo, suas competências específicas

permitem definir, em conjunto aos profissionais de tecnologia e da informação, requisitos de sistema vinculados à necessidade que motiva a produção dos documentos (ou a coleta de dados) para análise da informação.

Na produção documental, o arquivista irá colaborar nas especificações para criação dos documentos, por exemplo: se em valores numéricos, caracteres de texto, datas, som, imagem, quais formatos dos arquivos permitidos, quais limites de tamanho de arquivo, qualidade e demais elementos que serão úteis tanto para a preservação a longo prazo quanto para localização futura da informação. Ademais, como essas informações se relacionam com as demais, se organizam e se vinculam e como manter essas variáveis. Além disso, indicam os perfis de usuário que podem ter acesso, alterar, excluir, incluir informações. Também irão observar e colaborar para alertar quais informações podem ser alvo de quebra de privacidade ou possuem conteúdo sigiloso.

Na instituição onde atua, o arquivista deve ainda guardar os registros gerados na concepção dos sistemas, de modo a manter documentado o traçado e as decisões efetuadas para sua a parametrização, caso precise ser consultado, melhorado ou refeito pela mesma equipe ou outra que a substitua ao longo do tempo.

Ainda, geralmente o arquivista é o empregado designado como responsável na instituição pela gestão ao longo prazo, para planejar e orientar os aspectos para a preservação digital. Quando não há essa função designada os dados usualmente são perdidos, como ressaltado nos artigos objeto de revisão pelos autores Lemieux; Gormly; Rowledge (2014, p. 129 e 138) e Mcdonald e Léveillé, (2014, p. 107, 109 e 116).

Diante do exposto, apresentamos uma síntese do papel do arquivista, feita pelos autores Lemieux, Gormly e Rowledge:

> O gestor de documentos, sendo um especialista em dados, poderia atuar como parceiro no processo analítico, fornecendo informações sobre a localização dos dados e melhorando a compreensão e a confiança dos dados do analista visual, explicando seu contexto de criação, a história de sua estrutura e semântica e sua cadeia de custódia (…). Para aumentar a conscientização sobre sua especialização, os gestores de documentos precisarão se envolver com analistas visuais, cientistas de dados, analistas de negócios e outros dentro de organizações preocupadas com a análise e

> uso de Big Data[32] (LEMIEUX; GORMLY; ROWLEDGE, 2014, p. 139, tradução nossa).

Essa variedade de conhecimentos permite o arquivista estabelecer o valor das informações e identificar as relações entre si, dos documentos e das várias fontes armazenadoras de dados, habilidades essenciais ao cenário de utilização do *Big Data* o qual buscar valorar dados e atribuir-lhes novos usos.

[32] The records manager, as data expert, could function as a partner in the analytic process, providing information about data's location, and improving the visual analyst's understanding and trust of data through explaining their context of creation, the history of their structure and semantics and their chain of custody (…). To raise awareness of their expertise, records managers will need to engage with visual analysts, data scientists, business analysts and others within organisations concerned with the analysis and use of Big Data (LEMIEUX; GORMLY; ROWLEDGE, 2014, p. 139).

5 UMA APLICAÇÃO DO *BIG DATA* NA SECRETARIA DE INSPEÇÃO DO TRABALHO (SIT) NO MINISTÉRIO DO TRABALHO (MTB)

A fim de observar os conceitos trazidos por esta pesquisa e verificar na prática sua aplicação, este capítulo trata da utilização de uma ferramenta de *Big Data* em um projeto de pesquisa que envolve equipes de pesquisadores da Universidade de Brasília (UnB) das áreas de Ciência da Informação e de Tecnologia da Informação e, equipe de membros técnicos da Secretaria de Inspeção do Trabalho (SIT) do Ministério do Trabalho (MTB).

Em um turno de visitação, autorizada pelos professores coordenadores, foi observada uma apresentação do projeto em curso, com explicações quanto às necessidades de informação da Secretaria e os aspectos envolvidos na concepção, no desenvolver e na aplicação prática da ferramenta de *Big Data*.

O projeto tem por objetivo integrar e cruzar os dados de atividades-fim dos sistemas informatizados utilizados pelos Auditores-Fiscais do Trabalho em sua atividade de inspeção do trabalho e na fiscalização do Fundo de Garantia do Tempo de Serviço (FGTS). Os dados correlacionados e sintetizados trazem melhoria e qualidade à gestão da informação na Secretaria e, consequentemente, auxiliam a tomada de decisão dos auditores em suas tarefas de modo a proporcionar eficiência às auditorias realizadas junto aos empregadores em todo Brasil.

Em sua atuação, os auditores precisam realizar consultas frequentes e verificar dados para fazerem suas análises em vários sistemas, inclusive de outros órgãos, entre eles: Relação Anual de Informações Sociais (RAIS), Cadastro Geral de Empregados e Desempregados (CAGED), Seguro-Desemprego, Sistema Empresa de Recolhimento do FGTS e Informações à Previdência Social (SEFIP) e à base de dados da Receita Federal. São utilizadas também informações disponibilizadas pela Caixa Econômica Federal, referentes a recolhimentos de FGTS relativos a toda a base de empregadores, o sistema Fundo Garantidor de Crédito (FGC) para alimentação de sistema institucional (IDEB) e o arquivo SADO, que não está *online* e é atualizado e enviado pela Caixa. Estes sistemas são utilizados para planejamento fiscal e, também, como fonte de consulta dos auditores no momento da fiscalização.

Até o final de 2018, há previsão, pelo Ministério do Trabalho, da operação do e-Social em todas as empresas do Brasil. Esta será mais uma fonte de consulta dos

auditores, pois irá armazenar os eventos trabalhistas, inclusive as folhas de pagamento dos empregados, documento essencial para a fiscalização do FGTS.

A consulta feita aos dados, sistema a sistema, era morosa e as informações acabavam duplicadas, desencontradas ou desatualizadas dificultando as análises dos auditores. Além disso, acarretavam um custo muito alto, com muito esforço e desperdício de recursos materiais e financeiros no planejamento e execução da fiscalização do trabalho. Portanto, uma ferramenta capaz de integrar e cruzar os dados dos vários sistemas que possibilita gerar indicadores, planejar as ações de fiscalização e dar acesso às informações aos auditores em campo traz eficiência e, além disso, possibilita o combate às fraudes e à sonegação de impostos.

É possível notar que as bases de dados são várias e gigantescas, com complexidade entre os dados sobre todos os trabalhadores, empresas e relações de emprego do país, e, ainda, são constantemente atualizadas, o que torna necessária análises em tempo real das situações investigadas. Estes elementos são característicos de um cenário para utilização de um *Big Data* (dados em variedade, velocidade e volume) como visto na revisão de literatura.

Dessa forma, foi assim definida entre os membros do projeto e os auditores da Secretaria o desenvolvimento de uma ferramenta de *Big Data* para tratamento e organização de dados, permitindo o seu processamento e sua manipulação facilitada, em detrimento de outras opções tecnológicas que não abarcam essas características e não conseguiriam processar todo o volume do cenário em questão.

5.1 Estrutura do projeto

Conforme descrito por Araújo Jr. e Sousa (2016), os vários sistemas compõem um ecossistema de *Big Data*, com características e inter-relacionamento entre as partes que o compõem, seja na troca, localização ou seja na comparação entre os dados nas diferentes bases.

Para desenvolver esse ecossistema foi preciso planejar o *hardware* adequado e dimensioná-lo com foco tanto no processamento, quanto no armazenamento crescente de dados ao longo do tempo.

Para a economicidade do projeto ferramentas *software* livre[33] foram utilizadas, tais como: *Apache Hadoop, Apache Hadoop HDFS* e *Apache Hadoop MapReduce*.

Outro *software* utilizado, o produto *Hortonworks Data Platform* (HDP) inclui o *Apache Hadoop,* usado para armazenar, processar e analisar os dados em tratamento pela ferramenta de *Big Data*. A estrutura dos *hardwares* e *softwares* foi projetada para lidar com arquivos distribuídos *Hadoop, MapReduce, Pig, Hive, HBase* e *Zookeeper* e componentes adicionais.

O *software* para carga, visualização e análise de dados utilizado no projeto é o *Pentaho Business Analytics*[34]*,* que já cria *dashboards*[35], emite relatórios interativos, formas de visualização, análise e análise preditiva dos dados e, permite a manipulação e a utilização direta dos dados pelos usuários, no caso, os auditores.

5.2 Organização dos dados

Outros desafios significativos para iniciativas de *Big Data, a*lém da estruturação de *hardwares* e *softwares*, é determinar e tratar quais os dados que serão inseridos e/ou coletas pela ferramenta de maneira a se obter as informações desejadas.

Como visto na revisão de literatura, o *Big Data* está relacionado aos processos de negócio e essa definição dos processos e padrões que irão localizar e coletar os dados são a parte vital da ferramenta. Essa definição deriva da organização e da representação da informação, alvo da pesquisa dos profissionais da Ciência da Informação, ao estabelecer a modelagem e a arquitetura da informação a serem adotadas.

Na apresentação do projeto não foi mencionada a relação com a gestão de documentos, porém, compreende-se que ela estrutura previamente a organização

---

[33] *Software* livre é uma expressão utilizada para designar qualquer programa de computador que pode ser executado, copiado, modificado e redistribuído pelos usuários gratuitamente. Permite livre acesso ao código-fonte do *software* e a fazer alterações conforme as necessidades. Adaptado de: O que é Software livre?. Disponível em: https://www.significados.com.br/software-livre/. Acesso em: 25 de julho de 2018.

[34] É uma plataforma de análise de negócios que reúne insumos de TI e os usuários de negócios para facilitar o acesso, a integração, visualização e exploração de dados. Pentaho inclui descoberta de dados, integração de dados e análise preditiva.

[35] *Dashboards* são painéis nos sistemas de informação que mostram métricas e indicadores importantes para alcançar objetivos e metas traçadas de forma visual, facilitando a compreensão das informações geradas. Adaptado de: O que é *dashboard*?, por Rodrigo Nascimento, 19 de maio de 2017. Disponível em: http://marketingpordados.com/analise-de-dados/o-que-e-dashboard-%F0%9F%93%8A/. Acesso em: 25 de julho de 2018.

da informação, que, junto a representação, irão permitir que os grandes volumes de dados sejam gerenciados e manipulados, de maneira a garantir que os resultados sejam íntegros e úteis.

Devem ser levadas em consideração ainda na concepção da ferramenta a segurança e a proteção da privacidade dos dados coletados e operados.

É a partir da Especificação de Requisitos de Informação (ERI), que os usuários da ferramenta são consultados quanto as suas necessidades informacionais. Em reuniões de planejamento do projeto, a equipe dos pesquisadores junto aos auditores (usuários finais), levantaram as informações alvo de coleta e de interesse para as atividades que irão contribuir para o aprimoramento dos processos gerenciais sistêmicos na SIT.

Aos pesquisadores coube alinhar conceitos com os usuários finais (Auditores-Fiscais do Trabalho), compreender e traduzir as demandas em metadados para recuperação da informação, identificar os padrões para a definição de como a ferramenta opera (regras de negócio) para atender a demanda solicitada, a *interface* a ser apresentada ao usuário, o formato de apresentação dos dados, que filtros poderão ser aplicados para análises, determinar o funcionamento da localização, recuperação e comparação entre os vários tipos de dados pela ferramenta, dentre outras operações que são executadas.

A partir dos alinhamentos entre as equipes do projeto foram definidos índices para a modelagem da ferramenta, que relacionam a fonte da informação, sua localização dentro das bases disponíveis, a informação em si que será coletada e apresentada ao usuário. Esses índices irão transformar os dados em informação relevante para a auditoria, o que se denominou de Malhas de Fiscalização e Auditoria[36].

As malhas representam os elementos-alvo que irão nortear a fiscalização dos auditores em determinada apuração. Por exemplo, ao se definir a “ausência de controle de jornada”, traça-se a malha que extrai das bases de dados do e-Social, RAIS e CAGED, dados preenchidos nesses sistemas que comprovem a ausência de controle de jornada nos estabelecimentos fiscalizados, a ferramenta de *Big Data* trará como resultado quais foram as empresas que apresentaram inconsistências

[36] As Malhas de Fiscalização e Auditoria são extrações realizadas periodicamente nas bases de dados para identificar indícios de inconformidades na circularização das informações constantes nas bases de dados do *Big Data*.

nestas informações de controle de jornada, direcionando a atuação dos auditores a esses estabelecimentos em específico.

Outras malhas definidas como parâmetros para nortear análises dos auditores são: a inadimplência de salário, não concessão de férias anuais e pagamento fora do prazo legal. As inconsistências nessas informações apontadas pela ferramenta tornam-se alvos de fiscalização, que, por sua vez, orientam o início do trabalho dos auditores.

Nos controles das bases de dados executadas pela ferramenta de *Big Data* constam informações do tipo de disponibilização, local da disponibilização, nome da base de dados, periodicidade da disponibilização, origem do dado, o que possibilita a ferramenta atualizar a coleta dos dados com as informações de cada empresa em tempo real ou conforme forem modificadas pelo produtor do dado.

Dessa forma, os auditores possuem informações disponíveis de forma mais rápida e efetiva sem precisarem consultar manualmente cada base.

Outra facilidade trazida pela ferramenta é no trabalho em campo, em outra fase do projeto está sendo desenvolvido um aplicativo para dispositivos móveis, celulares ou *tablet* em que será possível o auditor no ato da fiscalização consultar informações, como por exemplo, ao se digitar o número do Cadastro Nacional da Pessoa Jurídica (CNPJ) da empresa no aplicativo, este traria resultados como seus dados completos, endereço (atrelando a localização ao *GPS*, caso necessário atualização no momento da fiscalização), quantidade de empregados ativos e a listagem de fiscalizações que já ocorreram na empresa, entre outros dados, proporcionando agilidade e conferência de informações aos auditores.

Pode-se observar que muitos dos dados são coletados a partir da análise de documentos de arquivo, produzidos digitalmente ou não, encontrados nas bases de dados mencionadas ou na própria inspeção dos auditores nos locais de trabalho, tais como: folhas de pagamentos, registro de ponto, livros fiscais e informações sobre as atividades laborais desenvolvidas (ARAÚJO JR.; SOUSA, 2016).

Neste contexto se verificam as relações existentes entre a ferramenta de *Big Data* e os documentos de arquivo, que irão comprovar os atos administrativos que produziram os dados. Dentre as relações estão a forma de interação entre as base de dados onde estão os documentos de arquivo e as bases dos dados no *Big Data*; definição de quais os dados necessários à fiscalização; a organização definida para os documentos que indicarão onde está o dado para coleta; a guarda qualificada da

informação pelo arquivo que mantém as características do documento de arquivo, bem como os requisitos para sua veracidade e valor, que tornará a informação útil a tomada de decisão.

Ainda, os documentos de arquivo que embasam a fiscalização, trazidos pelo *Big Data,* servem de subsídio para o auditor produzir novos documentos de arquivo, por exemplo, utilizando metadados estabelecidos previamente para a elaboração do relatório de fiscalização, que é um dos registros de sua atividade.

Com o exposto, percebe-se o papel desempenhado entre profissionais da ciência da informação, da tecnologia da informação junto a especialistas de uma determinada área, no caso, de auditoria do trabalho, para a qualidade da gestão da informação nas instituições, por meio da construção de uma ferramenta de *Big Data,* que proporciona a disponibilização de dados, sua maior precisão, que levam à melhores tomadas de decisão, maior eficiência operacional, redução de riscos e de custos.

6 CONSIDERAÇÕES FINAIS

A proliferação dos computadores, dos *smartphones* e do acesso facilitado à internet aumentou significativamente a capacidade da sociedade gerar, reunir, trocar, recuperar e examinar e utilizar dados com finalidades variadas. Soma-se a isso, a vantagem do acesso à informação à distância. A tecnologia proporciona inovações, gera conhecimentos e transforma a informação em um recurso cada vez mais valioso.

As análises de dados feitas em um *Big Data* proporcionam mais uma forma de obter vantagem competitiva, onde as novas formas de cruzamento de dados permitem desde prever e antecipar cenários, planejar vendas a até mesmo estratégias para cuidados de saúde de uma população ou decisão de um melhor trajeto a ser seguido em uma frota de caminhões. Para seu funcionamento é necessário investir em infraestrutura tecnológica e ferramentas para coleta e análise de dados.

Como em toda profissão, novos desafios surgem para os arquivistas e a Arquivologia na “era da informação”. Entre as fontes de dados em uma instituição, os documentos de arquivo devem ser devidamente geridos, com diretrizes e políticas para a sua produção, uso e preservação. Surgem então os questionamentos motivadores desta pesquisa de como os documentos arquivísticos serão aproveitados pelo *Big Data* como fonte de informação, ou de que maneira essa tecnologia irá se relacionar com os documentos de arquivo nas fases do ciclo vital.

Esta pesquisa se relaciona com esses elementos, pois discute a explosão informacional gerada com as tecnologias digitais, as diferentes formas de produção de dados e informações, bem como com a comunicação e recuperação da informação em um cenário de *Big Data* e busca possíveis interlocuções com a Arquivologia. Afinal, os documentos digitais são um novo desafio para o cenário tradicional da disseminação da informação (FROHMANN, 2008).

Demonstrou-se que o *Big Data* e a Arquivologia se interligam neste estudo, na organização, na busca pelo valor das informações, no auxílio à tomada de decisão, no acesso e difusão da informação. Os documentos de arquivo são reflexos das atividades institucionais e sendo autênticos servem de subsídios confiáveis para análises entre demais dados e informações. Assim sendo, o controle dos

documentos e, consequentemente, das informações que tramitam e são mantidas por meio da organização proporcionam o valor para a informação institucional que pode servir de análise em um *Big Data*. Este por sua vez, deve estar baseado nos processo de negócio, a fim de manter a organicidade dos documentos que irá utilizar como insumo e ser aplicável aos documentos de arquivo.

Em referência aos objetivos desta pesquisa considera-se que o *Big Data* é aplicável aos arquivos e ao documento de arquivo, no que tange ao uso como mais uma fonte de informação na coleta e no processamento de dados. Não se trata de uma tecnologia, ou técnica a ser adaptada aos documentos, ou ajustada à gestão documental ou vice-versa.

A gestão documental é que é capaz de preservar as características e especificidades dos documentos de arquivo e, como resultado, um acervo tratado e organizado poderá ser usado por um *Big Data*, que se favorece das características dos documentos de arquivo. Estas agregam qualidade à informação, veracidade e autenticidade aos dados coletados.

Pode-se verificar que a utilização dos documentos de arquivo no *Big Data* é mais uma forma para seu acesso, disseminação e evidencia a importância dos documentos na tomada de decisão.

Ademais, foi ainda apresentada uma descrição do desenvolvimento de uma ferramenta de *Big Data* em projeto no Ministério do Trabalho, onde a união entre profissionais da informação e da tecnologia propiciou a criação de uma ferramenta útil às necessidades das atividades dos Auditores Fiscais do Trabalho, com base na coleta, processamento e análise de diversos sistemas que continham dados e documentos de arquivo relativos a atividades-fim relacionadas à atividade de auditoria.

Os arquivistas continuam a desempenhar um papel decisivo na gestão de documentos e devem buscar capacitação para gerenciar, organizar e assegurar as características dos documentos de arquivo, sem abandonar os princípios e técnicas da área, seja no ambiente digital, seja em uma próxima inovação tecnológica. São essas informações que subsidiarão análises e tomadas de decisão e o valor agregado a elas que as definirão relevantes ou não.

Podemos observar, ainda, que no cenário de *Big Data,* assim como em um arquivo, nada se faz sem que a informação esteja minimamente organizada. Na possibilidade de uma recuperação futura de uma informação, ela deve ter sido

adequadamente produzida. Isto justifica o planejamento e a organização da criação dos documentos e de quais informações devem ser registradas e coletadas, para que elas existam e possam ser recuperadas. Em conformidade com os autores Lemieux, Gormly e Rowledge (2014), constata-se que:

> Embora este estudo seja de natureza exploratória e preliminar, os resultados indicam que o conhecimento e as habilidades dos profissionais de gestão documental permanecem relevantes e mais necessários do que nunca na era do *Big Data*[37] (LEMIEUX; GORMLY; ROWLEDGE, 2014, p. 139, tradução nossa).

Cabe à Arquivística continuar a aprofundar a reflexão sobre seus métodos, tendo em vista a realidade digital que se impõe às organizações, a crescente necessidade da execução de procedimentos de forma mais célere, o aumento do volume de dados produzidos e utilizados, bem como o constante desenvolvimento e atualização de ferramentas de trabalho e modernização da gestão.

6.1 Proposta para estudos futuros

Retomando a primeira fase exploratória desta pesquisa foram encontrados três artigos e um livro de produção nacional relacionados ao tema em análise e, também, com aspectos da Ciência da Informação. No quadro 6, estão relacionadas as referências dessas publicações. Entretanto, de acordo com a metodologia empregada e os critérios de pesquisa definidos não foram incluídos nos resultados da pesquisa bibliográfica, mas são aqui mencionados como recomendação para próximos estudos no tema:

- MIRANDA, M. L. C.; COSTA, Luciana S. *BIG DATA* não é uma tecnologia. **Data Grama Zero** - Revista de Informação – Coluna, v.15, n.3, jun/2014.

- PIMENTA, Ricardo M.. Big Data e controle da informação na Era digital: tecnogênese de uma memória a serviço do mercado e do estado. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 6, p. 7-24, 2013.

- RIBEIRO, C. J. S. Big Data: os novos desafios para o profissional da informação. **Informação & Tecnologia**, v. 1, p. 96-105, 2014.

[37] Although this study is exploratory and preliminary in nature, the findings indicate that the knowledge and skills of records professionals remain relevant and more necessary than ever in the era of Big Data (LEMIEUX; GORMLY; ROWLEDGE, 2014, p. 139).

- TARAPANOFF, K. (Org.). **Análise da informação para tomada de decisão**: desafios e soluções. Curitiba: InterSaberes, 2015. 365p.

Ademais, de forma a aprimorar os resultados desta pesquisa sugerem-se temas que podem ser abordados em análises posteriores:

- Aprofundar pesquisas de *Big Data* no âmbito da Ciência da informação;
- Relações entre outras tecnologias e a Arquivística;
- Ampliar o escopo desta pesquisa com foco em publicações produzidas no Brasil e/ou em países de língua espanhola;
- Ampliar os resultados das publicações explorando outras fontes de informação;
- Estudos para evidenciar o papel e a presença do arquivista junto a outros profissionais na implementação de tecnologias que lidem ou impactem os documentos de arquivo;
- Investigar o uso dos dados pessoais e a relação com o aspecto humano do *Big Data*.

REFERÊNCIAS

AGRAWAL, D. Analytics based decision making. **Journal of Indian Business Research**, 6(4), p. 332-340, 2014.

ALBUQUERQUE, A. C.; MADIO, T. C. C. A noção de classificação na arquivologia, biblioteconomia e museologia: abordagens teóricas. **Encontro Nacional de Pesquisa em Ciência da Informação**, v. 14, 2013.

ARAÚJO JÚNIOR, Rogério Henrique; SOUSA, Renato Tarciso Barbosa de. Estudo do ecossistema de *Big Data* para conciliação das demandas de acesso, por meio da representação e organização da informação. **Revista Ciência da Informação**, Brasília-DF, v.45 n.3, p. 187-198, set./dez. 2016.

BARTALO, Linete; APARECIDA, Nádina. **Gestão em Arquivologia**: abordagens múltiplas. Londrina: EDUEL, 2008

BELLOTTO, Heloísa Liberalli. **Arquivos permanentes**: tratamento documental. 4ª ed. Rio de Janeiro: Fundação Getúlio Vargas, 2006

_____. O sentido dos arquivos. In**: I Ciclo de Palestras da Diretoria de Arquivos Institucionais – DIARQ**. Universidade Federal de Minas Gerais. Belo Horizonte, 2014. (comunicação oral). Disponível em: <https://www.ufmg.br/diarq/anexos/wfd_14012774465385cc06bbb48--fala_bellotto.pdf>. Acesso em: 27 de março de 2018.

BHATTACHERJEE, Anol. **Social Science Research**: Principles, Methods, and Practices. University of South Florida, Tampa Bay: Open Access Textbooks Collection. Book 3. 2012

BRASIL. Arquivo Nacional. **Dicionário brasileiro de terminologia arquivística**. Rio de Janeiro: Arquivo Nacional, 2005. Disponível em: <http://www.conarq.arquivonacional.gov.br/Media/publicacoes/dicionrio_de_terminologia_arquivstica.pdf>. Acessado em: 25 de agosto de 2017.

_____. Comitê Executivo de Governo Eletrônico. **e-PING – Padrões de Interoperabilidade de Governo Eletrônico** – Documento de Referência Versão 2017, Brasília, 41 p.

_____. Conselho Nacional de Arquivos (Conarq). Câmara Técnica de

Documentos eletrônicos. **Glossário CTDE-Brasil**. Versão 7.0. Rio de Janeiro, 2016.

_____. Conselho Nacional de Arquivos (Conarq). **Resolução nº 25, de 27 de abril de 2007**. Dispõe sobre a adoção do Modelo de Requisitos para Sistemas Informatizados de Gestão Arquivística de Documentos - e-ARQ Brasil pelos órgãos e entidades integrantes do Sistema Nacional de Arquivos - SINAR. Rio de Janeiro, 2011.

_____. Lei nº12.527, de 18 de novembro de 2011. Regula o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei nº 8.112, de 11 de dezembro de 1990; revoga a Lei n. 11.111, de 5 de maio de 2005, e dispositivos da Lei no 8.159, de 8 de janeiro de 1991; e dá outras providências. **Planalto**. Brasília, 2011b. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>. Acesso em: 25 de julho de 2018.

_____. Ministério do Planejamento, Orçamento e Gestão. Secretaria de Logística e Tecnologia da Informação. e-MAG: Modelo de Acessibilidade em E-gov. **Ministério do Planejamento, Orçamento e Gestão, Secretaria de Logística e Tecnologia da Informação**. 2014. Brasília: MP, SLTI, 92 p.

CUKIER, K., MAYER-SCHOENBERGER, V. **Big Data**: como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana. Rio de Janeiro: Elsevier, 2013. 1 ed.

DAVENPORT, Thomas H. **Big Data no trabalho**: derrubando mitos e descobrindo oportunidades; trad. Cristina Yamagami. 1. ed., Rio de Janeiro: Alta Books, 2017.

DAVENPORT, T. H.; Barth, P.; Bean, R. How “big data” is different. Massachusetts Institute of Technology. **Sloan Management Review**, 54, 2012

DEMCHENKO Y.; NGO C.; DE LAAT C.; MEMBREY P.; GORDIJENKO D. Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure. In: Jonker W., Petković M. (eds.) **Secure Data Management**. Vol. 8425, p. 76-94. 2013. Disponível em: https://dx.doi.org/10.1007/978-3-319-06811-4_13. Acessado em: 27 de março de 2018.

DOLLAR, C. M. O impacto das tecnologias de informação sobre os princípios e práticas de arquivos: algumas considerações. **Revista do Arquivo Nacional**, v.7, n.1/2, jan./dez. 1994.

DURANTI, Luciana. Registros documentais contemporâneos como provas de ação. **Estudos Históricos**, Rio de Janeiro, v.7, n.13, p.49-64. 1994a.

FLORES, Daniel. **Manutenção da autenticidade, confiabilidade e fonte de prova dos documentos arquivísticos digitais (do SIGAD ao RDC-Arq)**. Câmara Municipal de São Paulo. São Paulo - SP. 124 slides, color, Padrão Slides Google Drive/Docs 4x3. Material elaborado para a Palestra na Unicamp, 19 de abril de 2016. Disponível em: <http://documentosdigitais.blogspot.com>. Acesso em: 09 de agosto de 2017.

FROHMANN, B. **O caráter social, material e público da informação**. In: FUJITA, M.; MARTELETO, R.; LARA, M. (Org.). A dimensão epistemológica da ciência da informação e suas interfaces técnicas, políticas e institucionais nos processos de produção, acesso e disseminação da informação. São Paulo: Cultura Acadêmica; Marília: Fundepe, 2008, p. 19-34. Disponível em: http://repositorios.questoesemrede.uff.br/repositorios/handle/123456789/829. Acesso em: 13 de abril 2018.

FURLAN, Patricia Kuzmenko; LAURINDO, Fernando José Barbin. Agrupamentos epistemológicos de artigos publicados sobre *big data analytics*. **Transinformação** [*online*]. 2017, vol.29, n.1, p. 91-100. Disponível em: http://dx.doi.org/10.1590/2318-08892017000100009. Acesso em: 13 de abril 2018.

GANTZ, John; REINSEL, David. Extracting Value from Chaos. **IDC IVIEW** - EMC Corporation. June, 2011. Disponível em https://uk.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf. Acesso em: 13 de abril 2018.

GOLDMAN, A.; KON, F.; PEREIRA JUNIOR, F.; POLATO, I.; PEREIRA, R. Apache Hadoop: Conceitos teóricos e práticos, evolução e novas possibilidades. **XXXI Jornadas de atualizações em informática (JAI)**, 2012.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO), *TC 46/SC 11. **ISO 15489-1:2001** Information and Documentation**.** Records Management. Part 1: General*, 1st ed., International Organization for Standardization, Geneva.

INTERNATIONAL COUNCIL ON ARCHIVES. (ICA). **Documentos de arquivo eletrônicos: manual para arquivistas.** Lisboa: Torre do Tombo, D.L. 2005. (ICA, Estudo n.º 16). Trad. de Electronic records: a work book for archivists. Paris: ICA, 2005. Disponível em: http://www.adporto.pt/ficheiros_a_descarregar/ica_estudo16_pt_4.pdf Acesso em: 13 de fevereiro 2018.

INTERPARES 2 PROJECT. **Diretrizes do produtor. A elaboração e a manutenção de materiais digitais**: diretrizes para indivíduos. TEAM Brasil. Trad. Arquivo Nacional e Câmara dos Deputados. 2002- 2007b. Disponível em: http://www.interpares.org/ip2/display_file.cfm?doc=ip2_creator_guidelines_booklet--portuguese.pdf. Acesso em: 31 de outubro de 2017.

JARDIM, J. M. **O conceito e a prática da gestão de documentos**. Acervo, Rio de Janeiro, v. 2, n. 2, p. 35-42, 1987.

_____. **As novas tecnologias da informação e o futuro dos arquivos**. Revista Estudos Históricos. Rio de Janeiro, v. 5, n. 10, p. 251-260, 1992.

_____. **A invenção da memória nos arquivos públicos**. Ciência da Informação, v. 25, n. 2, p. 1-13, 1996. Disponível em: <http://www.brapci.inf.br/v/a/860>. Acesso em: 25 de maio de 2018.

JARLBRINK, Johan; SNICKARS, Pelle. Cultural heritage as digital noise: nineteenth century newspapers in the digital archive. **Journal of Documentation**, Vol. 73 Issue 6, p. 1228-1243, 2017. Disponível em: http://dx.doi.org/10.1108/JD-09-2016-0106. Acesso em: 14 de janeiro de 2018.

LE COADIC, Yves-François**. A Ciência da Informação**. 2. ed. rev. e atual. Brasília, DF: Briquet de Lemos/Livros, 2004.

LEMIEUX, Victoria Louise; GORMLY, Brianna; ROWLEDGE, Lyse. Meeting Big Data challenges with visual analytics. **Records Management Journal**, Vol. 24, Issue 2, p. 122-141, 2014. Disponível em: http://dx.doi.org/10.1108/RMJ-01-2014-0009. Acesso em: 14 de janeiro de 2018.

LOPES, Luís Carlos. **A gestão da informação: as organizações, os arquivos e a informática aplicada**. Rio de Janeiro: Arquivo Público do RJ, 1997.

LOPEZ, André Porto Ancona. Princípios arquivísticos e documentos digitais. **Arquivo Rio Claro**, Rio Claro, n. 2, p. 70-85, 2004.

_____. **Sobre autenticidade e veracidade**. Brasília, 2010. Disponível em: http://diplomaticaetipologia.blogspot.com.br/2010/04/sobre-autenticidade-e-veracidade.html. Acesso em: 25 de dezembro de 2017.

LOUKISSAS, Yanni Alexander. Taking Big Data apart: local readings of composite media collections. **Information, Communication & Society**, 2016. Disponível em: http://dx.doi.org/10.1080/1369118X.2016.1211722. Acesso em: 14 de janeiro de 2018.

MACDONALD, S.; HEADLAM, N.. **Research Methods Handbook**: Introductory guide to research methods for social research. Manchester: Centre for Local Economic Strategies (CLES), 2009.

MAI, Jens-Erik. Big data privacy: The datafication of personal information. **The Information Society**, 32:3, p. 192-199, 2016. Disponível em: http://dx.doi.org/10.1080/01972243.2016.1153010. Acesso em: 14 de janeiro de 2018.

MARCHIORI, P. Z. A ciência e a gestão da informação: compatibilidades no espaço profissional. **Ciência da Informação**, Brasília, v. 31, n. 2, p. 72-79, maio/ago. 2002.

MAYER-SCHÖNBERGER, Viktor; CUKIER, Kenneth. ***BIG DATA***: como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana; tradução Paulo Polzonoff Junior. 1º ed., Rio de Janeiro: Elsevier, 2013.

MCAFEE, Andrew; BRYNJOLFSSON, Erik. Big Data: The Management Revolution. **Harvard Business Review**, October 2012. Disponível em: https://hbr.org/2012/10/big-data-the-management-revolution. Acesso em: 08 de agosto de 2017.

MCDONALD, John; LÉVEILLÉ, Valerie. Whither the retention schedule in the era of big data and open data?. ***Records Management Journal****,* Vol. 24 Issue: 2, 2014. p. 99-121. Disponível em: https://dx.doi.org/10.1108/RMJ-01-2014-0010. Acesso em: 14 de janeiro de 2018.

MC GEE, J.; PRUSAK, L. **Gerenciamento Estratégico da Informação**: aumente a competitividade e a eficiência de sua empresa utilizando a informação como uma ferramenta estratégica. Tradução de Astrid Beatriz Figueiredo. Rio de Janeiro: Campus, 1994.

MIRANDA, M. L. C.; COSTA, Luciana S. *BIG DATA* não é uma tecnologia. **Data Grama Zero** - Revista de Informação – Coluna, v.15, n.3, jun/14.

MORENO, Nádina Aparecida. A Informação Arquivística e o Processo de Tomada de Decisão. **Informação & Sociedade**: estudos, João Pessoa, v.17, n.1, p. 13-21, jan 2007.

OLIVEIRA, Carolina. **Um estudo de caso sobre *datasets* do Ministério da Justiça**: dados brutos ou documentos arquivísticos?. 2015. Dissertação (Mestrado Profissional em Gestão de Documentos e Arquivos) - Universidade Federal do Estado do Rio de Janeiro.

PIMENTA, Ricardo M.. *Big Data* e controle da informação na Era digital: tecnogênese de uma memória a serviço do mercado e do estado. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 6, p. 7-24, 2013.

REGO, Bergson Lopes. **Gestão e Governança de Dados**: Promovendo dados como ativo de valor nas empresas. Brasport, 2013.

RIBEIRO, C. J. S. Big Data: os novos desafios para o profissional da informação. **Informação & Tecnologia**, v. 1, p. 96-105, 2014.

ROCHA, C. L.; SILVA, M. da. Padrões para garantir a preservação e o acesso aos documentos digitais. Acervo, Rio de Janeiro, v. 20, n. 1-2, p. 113-124, jan/dez 2007. Disponível em: <http://www.revistaacervo.an.gov.br/seer/index.php/info/article/view/142>. Acesso em: 25 de julho de 2018.

RODRIGUES, Ana Márcia Lutterbach. **A teoria dos arquivos e a gestão de documentos**. Revista Perspectiva em Ciência da Informação, Belo Horizonte, v.11 n.1, p. 102-117, jan./abr. 2006

RONDINELLI, Rosely Curi. **Gerenciamento arquivístico de documentos eletrônicos**: uma abordagem teórica da diplomática arquivística contemporânea. 4ª. Ed. – Rio de Janeiro: Editora FGV, 2005. 160 p.

_____. **O Conceito de documento arquivístico frente à realidade digital**: uma revisitação necessária / Rosely Curi Rondinelli -- 2011. 270 f.: il. Tese (Doutorado em Ciência da Informação) – Universidade Federal Fluminense, Programa de Pós-Graduação em Ciência da Informação, Instituto de Arte e Comunicação Social, Instituto Brasileiro em Ciência e Tecnologia, Niterói, 2011.

_____. **O documento arquivístico ante a realidade digital**: uma revisão conceitual necessária. 1ª. Ed. – Rio de Janeiro: Editora FGV, 2013. 280 p.

ROUSSEAU, Jean-Yves; COUTURE, Carol. **Os fundamentos da disciplina arquivística**. Lisboa: Publicações Dom Quixote, 1998.

RUBEL, Alan; JONES, Kyle M. L.. Student privacy in learning analytics: An information ethics perspective. **The Information Society**, 32:2, p. 143-159, 2016. Disponível em: http://dx.doi.org/10.1080/01972243.2016.1130502. Acesso em: 14 de janeiro de 2018.

SANT'ANA, Ricardo Cesar Gonçalves. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Informação & Informação,** Londrina, v. 21, n. 2, p. 116 – 142, maio/ago., 2016.

SANTOS, Henrique Machado dos; FLORES, Daniel. Repositórios digitais confiáveis para documentos arquivísticos: ponderações sobre a preservação em longo prazo. **Revista Perspectiva em Ciência da Informação [*online*].** 2015, vol.20, n.2, p. 198-218.

_____. O documento digital no contexto das funções arquivísticas. **Páginas a & b - arquivos & bibliotecas**, v. 1, p. 165-177, 2016.

SANTOS, Vanderlei Batista dos. **A prática arquivística em tempos de gestão do conhecimento**. In: SANTOS, Vanderlei Batista dos; INNARELLI, Humberto Celeste; SOUSA, Renato Tarciso Barbosa de (Org.). Arquivística: temas contemporâneos: classificação, preservação digital, gestão do conhecimento. 3. ed. Brasília, DF: SENAC, 2008.

_____. Preservação de documentos arquivísticos digitais. **Revista Ciência da Informação (*Online*)**, v. 41, p. 114-126, 2012.

SCHELLENBERG, Theodore Roosevelt. **Arquivos modernos**: princípios e técnicas. Trad. Nilza Teixeira Soares. 6 ed. Rio de Janeiro: Editora FGV, 2006.

SEMIDÃO, R. A. M. **Dados, Informação e Conhecimento enquanto elementos de compreensão do universo conceitual da Ciência da Informação**: contribuições teóricas. Marília, 2014. 198 f. Dissertação (Mestrado). Programa de Pós-Graduação em Ciência da Informação - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista – UNESP, Marília, 2014.

SILVEIRA, M.; MARCOLIN, C. B.; FREITAS, H. M. R. Uso Corporativo do *Big Data*: Uma Revisão de Literatura. **Revista de Gestão e Projetos**, v. 6, n. 3, p. 44-59, 2015.

SOUSA, Renato Tarciso Barbosa de; ARAÚJO JÚNIOR, Rogério Henrique de. A indexação e criação de taxonomias para documentos de arquivo: proposta para a expansão do acesso e integração das fontes de informação. ***Brazilian Journal of Information Science: Research Trends,*** v. 11, n. 4, p. 47-56, 2017.

TARAPANOFF, K. (Org.). **Análise da informação para tomada de decisão**: desafios e soluções. Curitiba: InterSaberes, 2015. 365p.

TAURION, Cezar. **BIG DATA**. Rio de Janeiro: Brasport Livros e Multimídia Ltda., 2015.

TAYLOR-SAKYI, Kevin. **Big Data**: Understanding Big Data. Birmingham, England, Aston University, 2016. Disponível em: https://arxiv.org/ftp/arxiv/papers/1601/1601.04602.pdf. Acesso em: 14 de janeiro de 2018.

ZIKOPOULOS, P.; LIGHTSTONE, S.; HURAS, M.; SACHEDINA, A. et al. (2013). New dynamic in memory analytics for the era of big data. **IBM Data Management Magazine**, (4), p. 1-47, 2013.